MAITRISE DES SCIENCES DU LANGAGE
MENTION INDUSTRIES DE LA LANGUE

MEMOIRE DE RECHERCHE

Le Traitement automatique et lexicographique des locutions verbales figées en français

Présenté par

Marie-Véronique LEROI

Sous la direction de Mr Fleury

UNIVERSITE PARIS III
SORBONNE NOUVELLE
ILPGA

Remerciements

Je tiens à remercier toutes les personnes qui ont pu m'aider et qui sont intervenus d'une manière ou d'une autre dans la réalisation de ce mémoire, notamment :

- Monsieur Fleury pour sa disponibilité, son écoute et son aide,

- Madame Samvelian pour ses précieux conseils,

- et enfin Monsieur Salem, qui m'a permis de recentrer le sujet de ce travail, pour son amabilité.

Je tiens également à remercier mes amis et amies qui m'ont beaucoup encouragée tout au long de cette année.

Je dédie ce mémoire à mes parents à qui je dois tout.

Table des Matières

Introduction

« Mot composé, locution, idiotisme, expression idiomatique, phraséologisme, cliché, proverbe, dicton, etc…, autant de termes, souvent mal définis, pour décrire l'extrême variété des expressions figées et consacrées par l'usage.»

Cette citation extraite de l'article de Georges Misri intitulé *« Approches du figement linguistique : critères et tendances »*, publié en 1987, montre l'ampleur du phénomène que constituent les expressions figées. Comme l'a démontré Maurice Gross, ce phénomène n'a rien d'exceptionnel et est même très courant dans les langues naturelles. Ces expressions figées sont désignées par un nombre important de dénominations usuelles et communes dans certains cas (locutions, expressions figées, noms composés, …), spécifiques dans d'autres cas (synapsie, synthème, …). Différents auteurs sont à l'origine de ces diverses dénominations. Cette abondance est due aux divers domaines d'études que touchent les expressions figées. Certains auteurs, en effet, proposent un traitement de nature strictement ou essentiellement syntaxique – c'est le cas de Maurice Gross -, d'autres auteurs estiment qu'il s'agit d'un phénomène d'ordre lexicologique dans la mesure où les expressions figées constituent des unités lexicales à proprement parler. Ivan Fonagy a traité les expressions figées dans un cadre d'études phonologique et discursif. Le domaine du TAL (Traitement Automatique des Langues) a consacré bon nombre d'études aux expressions figées. En effet, nombre de travaux du LADL (Laboratoire d'Automatique Documentaire et de Linguistique) et du CERIL (Centre d'Etudes et de Recherches en Informatique et Linguistique) ont traité les expressions figées dans une perspective de traitement automatique.

Parmi l'ensemble des séquences qui présentent un caractère figé, il est possible de remarquer que les noms composés et les locutions verbales sont les constructions qui ont fait l'objet du plus grand nombre d'études chez les linguistes.

Le travail présenté ici porte sur le traitement automatique et lexicographique des locutions verbales figées en français et procède donc à une description du figement pour ce faire. Ces locutions qui posent de nombreuses difficultés dans le domaine du Traitement Automatique des Langues (TAL) s'avèrent être un obstacle non négligeable pour les apprenants étrangers du français. Le principal problème posé par ce type d'expressions est leur reconnaissance dans un corpus donné. Un outil en traitement automatique doit être en mesure de reconnaître ces séquences pour fournir une analyse correcte du texte. Les critères permettant de distinguer les séquences libres des séquences figées varient autant que les différentes dénominations proposées par les auteurs. Les locutions verbales que nous allons étudier plus précisément dans ce travail, posent davantage de problèmes car il est difficile de les reconnaître de manière automatique dans la mesure où les verbes dans ces locutions connaissent les variations qui leur sont propres, à savoir la conjugaison, l'accord en genre et en nombre avec le sujet. Les structures spécifiques sont assez rares. Un autre problème qui se pose est celui de la discontinuité de ces locutions : en effet, certaines locutions verbales permettent l'insertion de modifieurs et sont donc discontinues. Nous verrons dans le cadre de ce travail que divers travaux menés précédemment proposent des méthodes spécifiques pour résoudre ces problèmes. Nous étudierons plus précisément le traitement proposé par deux

outils qui permettent d'analyser lexicalement et syntaxiquement de gros corpus et de procéder à la recherche d'information ou à l'extraction terminologique.

Ce projet de mémoire a conduit à la possibilité de création d'une application qui pourrait proposer un traitement pour ces expressions figées et en particulier pour les locutions verbales. Le projet a donc consisté en l'élaboration d'un logiciel présentant les caractéristiques d'un éditeur de texte classique au premier abord mais offrant des fonctionnalités spécifiques. Ce logiciel dénommé Verbalex a été conçu afin de répondre à certaines attentes quant au traitement automatique ou automatisé des locutions verbales figées en français et de souligner les difficultés que peuvent présenter la conception d'un outil à tout point de vue et plus particulièrement la conception d'un outil visant le traitement de séquences figées. L'application vise à extraire les locutions verbales figées du français à partir d'un corpus, pour cela nous procédons à l'extraction des expressions verbales et nous tentons de déterminer si les formes extraites sont figées ou non. Ces expressions verbales seront désignées dans le cadre de ce travail par le terme « locutions verbales ». Nous verrons que ce terme ne reçoit pas la même acception selon les auteurs et nous donnerons une définition qui correspond au mieux à la perspective d'étude que nous avons adoptée dans le cadre de ce travail. Une fois extraites, ces locutions constituent une base de données dictionnairique disponible dans l'application. Ce dictionnaire électronique serait enrichi manuellement en ce qui concerne les informations d'ordre morphosyntaxique ou sémantique sur les locutions.

Le logiciel permettrait aussi à l'utilisateur de compléter manuellement le dictionnaire électronique constitué à partir des corpus passés en traitement. Les différentes étapes de traitement des corpus ouverts dans l'application seront visibles sans altérer le fichier original. Les principaux traitements opérés seront la constitution d'un dictionnaire de formes graphiques, autrement dit le recensement de toutes les chaînes de caractères qui apparaissent dans le corpus qui pourront être affichées selon l'ordre alphabétique ou selon leur fréquence d'apparition. Nous verrons également que l'application permet de procéder à l'étiquetage et à la lemmatisation du fichier, étapes qui sont primordiales dans toute démarche de traitement textuel. Il sera également possible de procéder à des recherches dans les différents états du corpus.

Le présent mémoire de recherche va donc retracer les principes et les démarches qui ont précédé à la création de cet outil. Mais avant cela, nous procéderons à une étude détaillée du figement et des propriétés des locutions verbales.

Nous avons distingué deux parties dans le travail présenté dans les pages suivantes.

Nous étudierons dans la première partie le processus du figement et nous tenterons de définir et décrire ce phénomène. Nous déterminerons ensuite ce qu'est une locution verbale figée en français en étudiant les principaux traits qui la caractérisent. Pour cela, nous verrons les différentes acceptions proposées par les auteurs, et celles que nous adopterons ainsi que les différents critères de reconnaissance possibles au niveau linguistique.

Dans la deuxième partie, nous procéderons à une étude du figement et plus particulièrement des locutions verbales du point de vue du traitement automatique. Cette seconde partie est divisée en deux sous-parties. La première nous permettra d'étudier et de

décrire quels types de traitements sont possibles pour les expressions figées et quelles sont les difficultés que ce traitement soulève. Nous décrirons également différentes méthodes et outils qui proposent un traitement des expressions figées. Nous étudierons plus en avant deux outils qui permettent l'analyse de gros corpus : Intex et Unitex conçus respectivement par le LADL et l'Institut d'Electronique et d'Informatique Gaspard Monge. Nous procéderons également à une application concrète de l'utilisation de ces deux outils sur un corpus constitué de deux articles de l'édition électronique du Monde et étudierons les résultats produits.

La seconde sous-partie est consacrée à l'étude détaillée de l'élaboration de l'application Verbalex. Nous aborderons les différents principes et méthodes utilisées pour la création du logiciel ainsi que les divers problèmes rencontrés dus à l'implémentation du programme mais aussi aux contraintes liées au traitement des locutions verbales.

Partie A

Les propriétés linguistiques du figement et des locutions verbales

---

I / Les locutions verbales figées en français

Nous allons tenter dans cette partie de déterminer ce que désigne une « locution verbale figée ». Avant cela il nous faut définir ce qui est entendu par l'adjectif « figé » et donner une description du figement. Nous aborderons ensuite les locutions avec plus de précision et en particulier les locutions verbales figées.

1. Le Figement

1.1 Le figement : une exception dans la langue

Le figement est un phénomène linguistique complexe longtemps considéré comme irrégulier et qui a donc un caractère marginal dans la langue. Maurice Gross *(1985)* a pourtant démontré dans ses travaux que d'un point de vue statistique ces sentiments d'irrégularité et d'exception n'avaient pas lieu d'être. En effet, il existerait près de 1800 constructions verbales qui ne mettent pas en jeu un emploi spécifique du verbe. C'est le cas d'une phrase telle que :

Ex1.1.a) : « Luc lèche le plat ».

A contrario, 8000 constructions verbales seraient figées. L'exemple suivant montre bien que le verbe « lécher » n'est pas utilisé avec le même emploi que précédemment :

Ex1.1.b) : « Luc lèche les bottes de Max ».

M. Gross estime donc qu' « ignorer ces constructions revient à ignorer une bonne partie du langage ».

Otto Jespersen a été l'un des premiers linguistes à aborder ce phénomène qu'est le figement. Dans son ouvrage *« Philosophy of Grammar » (1924),* il distingue deux principes dans les langues : la liberté combinatoire et le figement. Cette manière d'aborder les langues confère un caractère essentiel au processus de figement.

Weinrich *(1969)* accordait aussi une grande importance aux expressions figées. Il disait à propos du figement : « Ce qui avait longtemps été considéré comme un phénomène marginal, comme une série d'exceptions, se révèle être en fait caractéristique des langues

humaines naturelles ». Gaston Gross (*1981*) surenchérit en accordant la même importance au phénomène des expressions figées qu'à la double articulation d'André Martinet (*1967).*

A l'inverse, certains auteurs ont tendance à accorder une trop grande importance au phénomène en disant que tout est phraséologique.

Les nombreuses et diverses définitions et dénominations qui ont été introduites par les différents auteurs et leurs ouvrages pour décrire ce même phénomène ont contribué à apporter au figement un caractère marginal et irrégulier.

1.2 Définitions

Les définitions proposées pour le nom « figement » ou l'adjectif « figé » sont très variées. L'adjectif « figé » est défini de la manière suivante dans divers dictionnaires et ouvrages:

Lexis : « Figé : se dit d'un mot, d'une construction qui cessent de subir dans la langue une évolution. »

Petit Robert : « Expression, locution figée : dont on ne peut changer les termes et qu'on analyse généralement mal ».

Ces définitions sont pour le moins laconiques et se contentent de souligner l'existence du phénomène tout en supposant que celui-ci est irrégulier. La remarque faite sur les expressions figées donnée par Alain Rey et Sophie Chantreau dans leur *« Dictionnaire d'expressions et locutions » (1997)* fournit davantage de précision :

Dictionnaire d'expressions et locutions (1997): « Un lexique ne se définit pas seulement par des mots simples et complexes, mais aussi par des suites de mots convenues, fixées, dont le sens n'est guère prévisible […]. Ces séquences, on les appelle en général des locutions ou des expressions. »

Les dictionnaires de linguistique se veulent plus précises et détaillées.
Le *Dictionnaire de linguistique (1973)* donne donc une définition un peu moins vague et s'appuie sur des exemples :

Dictionnaire de Linguistique (Larousse) (*1973*): « **Figement** : Le figement est un processus linguistique qui, d'un syntagme dont les éléments sont libres, fait un syntagme dont les éléments ne peuvent être dissociés. Ainsi, les mots composés (compte rendu, pomme de terre, etc…) sont des syntagmes figés. »

Nous verrons plus avant que la composition et le figement sont des phénomènes distincts et que tous les mots composés ne sont pas nécessairement figés.

Dictionnaire de Linguistique et des Sciences du Langage (*1994*): « **Figement** : Le figement est le processus par lequel un groupe de mots dont les éléments sont libres devient

une expression dont les éléments sont indissociables. Le figement se caractérise par la perte du sens propre des éléments constituant le groupe de mots, qui apparaît alors comme une nouvelle unité lexicale, autonome et à sens complet, indépendamment de ses composants. »

Ces différentes définitions tendent à montrer que le figement est un phénomène hors norme et irrégulier.

J.C. Anscombre (*1990*) définit le figement comme étant un processus au terme duquel le locuteur n'est plus capable de déterminer le sens d'une séquence à partir de celui de ses constituants.

Georges Misri (*1987*), quant à lui, désigne sous le terme de « figement » « tout groupe de monèmes qui présente un blocage total ou quasi total des axes paradigmatiques et syntagmatiques, c'est-à-dire une impossibilité ou une réduction importante des possibilités de commutation et / ou d'expansion partielle. »

Dans la section suivante, nous allons voir que les auteurs ont proposé différents termes pour décrire le figement. Ces divers termes permettent aux auteurs d'exprimer des nuances dans leur théorie du figement.

1.3 Une profusion terminologique

Certains auteurs classiques ont proposé différents termes pour décrire le figement : ces différentes dénominations illustrent en fait des points de vue théoriques divergents et permettent de mettre en relief le fait que le figement est un phénomène irrégulier.

- *Ferdinand de Saussure* a parlé dans le *« Cours de Linguistique Générale » (1916)* d' «expression ou de locution toute faite ». Cette qualification laisse transparaître le caractère immuable inhérent à ce type d'expressions :

«Le propre de la parole, c'est la liberté des combinaisons. Il faut donc se demander si tous les syntagmes sont également libres. On rencontre un grand nombre d'expressions qui appartiennent à la langue ; ce sont les locutions toutes faites, auxquelles l'usage interdit de rien changer […].» (Ferdinand de Saussure, *« Cours de Linguistique Générale », 1916*).

- *Charles Bally* dans son *« Traité de Stylistique » (1909)* consacre un chapitre aux locutions phraséologiques. Parmi ces locutions, il distingue les « séries » des « unités » phraséologiques. Les séries phraséologiques sont des locutions où la cohésion des termes est relative. C. Bally les définit comme suit :

« Les éléments du groupe conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche, de sorte que l'ensemble présente des contours arrêtés et donne l'impression du ‘‘ déjà vu''. »

Les unités phraséologiques désignent des locutions où la cohésion des termes est absolue. Une unité phraséologique est définie de la manière suivante par C. Bally :

« Une unité phraséologique représente un groupe de mots où «les mots qui composent le groupe perdent toute signification et l'ensemble seul en a un. […] Cette signification doit être nouvelle et non équivalente à la somme des significations des éléments. »

Il est possible de voir que le critère intuitif est privilégié par Bally pour la distinction des locutions phraséologiques.

- <u>*Henri Frei*</u>, dans la *« Grammaire des Fautes »*, publié en *1969*, parle quant à lui de brachysémie (ou figement). Ce terme est synonyme de brièveté sémantique :

« Le mécanisme de la brachysémie ou brièveté sémantique est le figement d'un syntagme, c'est-à-dire d'un agencement de deux ou plusieurs signes, en un signe simple. La brachysémie, brièveté sémantique se distingue de la brachylogie, brièveté formelle. »

- <u>*Emile Benveniste*</u> a proposé dans les ouvrages *« Problèmes de linguistique Générale »* et « *Formes nouvelles de la composition nominale » (1966)* distingue trois types de formes complexes qui sont représentées dans le tableau ci-dessous :

<u>**Termes**</u>	*Définitions*	exemples
1. **Composés**	*Unités à deux termes identifiables par le locuteur*	portefeuille
2. **Conglomérats**	*Unités nouvelles formées de syntagmes complexes comportant plus de deux éléments*	Va-nu-pieds Meurt-de-faim
3. **Synapsie**	*Groupe entier de lexèmes, reliés par divers procédés, et formant une désignation constante et spécifique*	Fusil de chasse

<u>*Tableau 1.3.a : Les trois différents types d'unités complexes selon E. Benveniste*</u>

E. Benveniste propose donc le terme de « synapsie » pour désigner des séquences de mots présentant un caractère figé. Selon sa définition une synapsie représenterait une unité de signification composée de plusieurs morphèmes lexicaux. Benveniste utilise ce terme pour mettre en évidence le fait qu'il s'agit d'un modèle de construction différent de celui de la composition classique. La synapsie est donc différente du mot composé et du mot dérivé comme le montrent les exemples ci-dessous :

<u>**Terme**</u>	exemples
Synapsie	« machine à coudre »

Mot composé	« timbre-poste »
Mot dérivé	« ferblanterie »

Tableau 1.3.b : Exemples proposés par Benveniste pour la distinction de la synapsie des autres mots construits.

- *André Martinet*, dans un article intitulé *« Syntagme et Synthème »*, paru en 1967, introduit le terme de « *synthème* ». Martinet se place du point de vue syntaxique fonctionnaliste pour définir le « synthème ». Ce terme désignerait donc les « unités linguistiques dont le comportement syntaxique est strictement identique à celui des monèmes avec lesquels ils commutent, mais qui peuvent être conçus comme formés d'éléments sémantiquement identifiables. D'après cette définition, le synthème représenterait donc une séquence formée de plusieurs monèmes lexicaux fonctionnant comme une unité syntaxique minimale. Les mots dérivés sont considérés comme des synthèmes.

- *Bernard Pottier* dans son ouvrage *« Linguistique Générale»*, et *« Introduction à l'étude des Structures grammaticales fondamentales »* (1962) utilise le terme de « lexie » pour désigner les unités lexicales. Une lexie est une unité lexicale mémorisée. Il distingue trois types de lexies.

Terme	*Exemples*
Lexie Simple	*Cheval*
Lexie Composée	*Cheval-vapeur*
Lexie Complexe	*Cheval marin*

Tableau 1.3.c : Les trois différents types d'unités lexicales selon B. Pottier

La lexie composée est un ensemble comprenant plusieurs mots intégrés ; la graphie et plus précisément le trait d'union permet de reconnaître une lexie composée.

La lexie complexe désigne « une séquence en voie de lexicalisation à des degrés divers. La lexie complexe est une séquence qui peut être figée ou non. Le critère de séparabilité permet de les reconnaître : en effet, il sera question d'une lexie complexe si les éléments du groupe ne sont pas séparables ; et à l'inverse il s'agira d'un syntagme si les éléments du groupe sont séparables.

- *Maurice Gross* (1985) parle de « phrases figées ». La phrase constitue dans ses travaux, qui s'inscrivent dans la théorie du lexique-grammaire, l'unité sémantique de base ; les mots ou les morphèmes ne sont donc pas dans le cadre de ses travaux les unités minimales.

Il n'est donc jamais question de « locution » pour désigner les séquences figées tout comme il n'est jamais question de « syntagme » pour se référer aux séquences libres.

- Gaston Gross a introduit la notion de figement et d'expressions figées dans *« Les expressions figées en français, noms composés et autres locutions »*, son ouvrage qui date de *1996*. Deux principes ont leur importance dans la reconnaissance des expressions figées : il s'agit de l'opacité sémantique et la liberté combinatoire. Nous reviendrons dans les sections suivantes sur ces deux principes.

1.4 Caractéristiques générales du figement

Les manuels classiques de lexicologie abordent le figement dans la partie traitant de la composition. Nous allons donc dans un premier temps, tenter de définir quels sont les liens entre le figement et la composition puis découvrir les caractéristiques générales des expressions figées.

1.4.1 Figement et Composition

Gaston Gross étaye sa description du figement en utilisant un certain nombre de termes et de définitions spécifiques qu'il est possible de retrouver chez d'autres auteurs:

- Un groupe ou un syntagme est dit « *libre* » s'il correspond à une « séquence générée par les règles combinatoires mettant en jeu à la fois des propriétés syntaxiques et sémantiques ». l'adjectif « libre » s'oppose donc à l'adjectif « figé ».

- Un *idiotisme* (gallicisme, anglicisme ou germanisme) est une séquence que l'on ne peut pas traduire terme à terme dans une autre langue.

- Un « *mot racine* » ou un « *mot simple* » désigne toute unité qui n'est susceptible d'aucune décomposition.

- Un mot qui n'est pas un mot simple est alors dit « *construit* ». Les mots construits sont donc des mots composés de différents morphèmes autonomes.

G. Gross fait ressortir deux types de mots construits :

1. Les mots *dérivés* que l'on obtient par l'affixation d'un préfixe ou d'un suffixe à une base donnée ;

2. Les mots *polylexicaux* (ou mots complexes) qui désignent « toute unité composée de deux ou plusieurs mots simples ou dérivés préexistants. » Ces mots peuvent être soudés et donc ne pas comporter de séparateurs.

Le schéma suivant pourrait représenter les différents types d'unités lexicales :

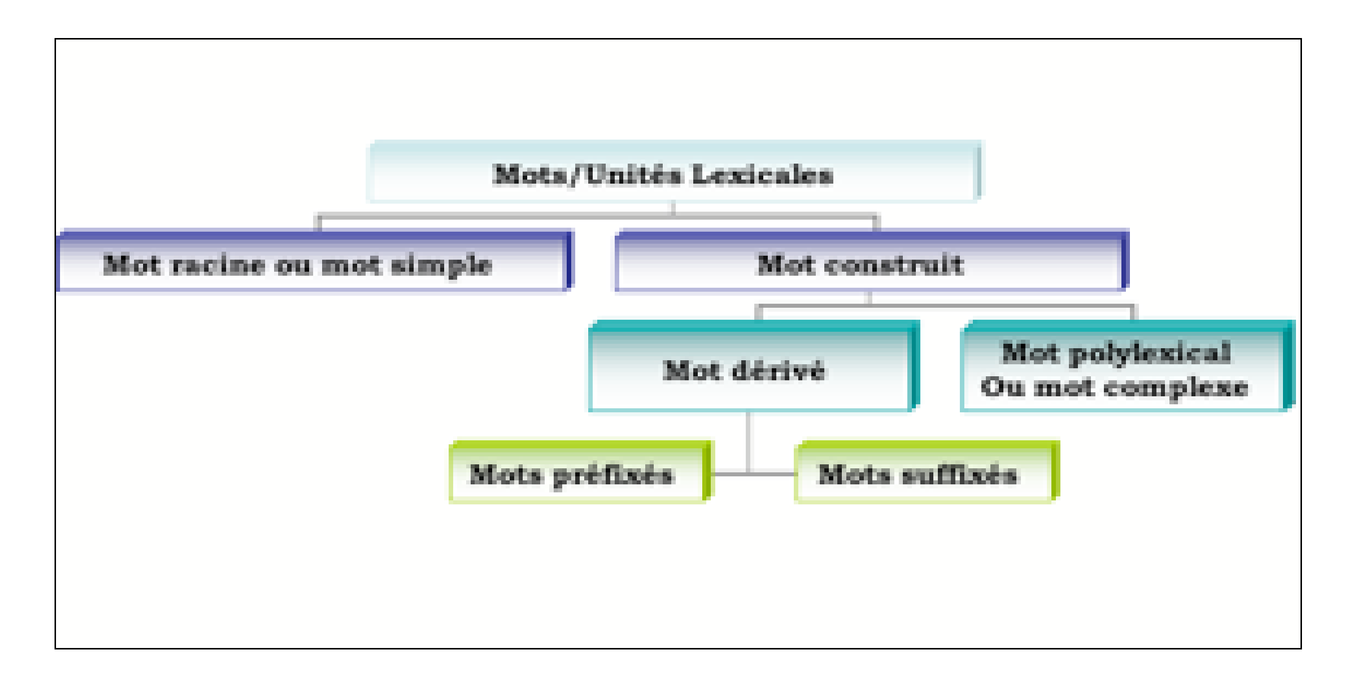


Figure 1.4.1.a) : Les différents types d'unités lexicales selon Gaston Gross

La composition comme le montre la *figure1.4.1.a)* ci-dessus est donc un des moyens de formation de nouvelles unités lexicales disponibles en français. La composition est traditionnellement opposée à la dérivation. La dérivation est un procédé récursif : la base à laquelle est affixée un préfixe ou un suffixe peut elle-même être un mot dérivé. La composition est donc moins productive que la dérivation quant à la formation des nouveaux mots en français.

Figement et composition sont souvent amalgamés et considérés comme des synonymes. Mais cela n'est pas avéré ; en effet, une suite composée n'est pas nécessairement figée. Les suites composées peuvent être sémantiquement transparentes et à cet égard ne seront donc pas considérées comme figées.

Gaston Gross met en évidence deux types de contraintes qui interviennent dans sa description du figement :

- une contrainte d'ordre *syntaxique* : une suite donnée est-elle syntaxiquement libre ?
- une contrainte d'ordre *sémantique* : l'opacité sémantique ; cette suite est-elle sémantiquement transparente ou opaque ?

D'après ces indications, une suite peut être considérée comme étant figée quand celle-ci n'est pas libre syntaxiquement et est sémantiquement opaque. Ces deux contraintes vont de pair.

Le terme « polylexical » utilisé dans la *figure 1.4.1.a)* peut se définir de la manière suivante : une suite est dite« polylexicale » quand elle est composée de plusieurs éléments lexicaux qui ne jouent pas de rôle extérieur à la séquence. Autrement dit les éléments lexicaux contribuent uniquement à la constitution de la suite.

1.4.2 Les critères de reconnaissance

- Des critères génériques

Dans l'ouvrage intitulé *« Introduction à la lexicologie, sémantique et morphologie »* (LEHMAN Alise et MARTIN-BERTHET Françoise, *2000*), il est possible de discerner trois critères qui permettent de distinguer les syntagmes libres des syntagmes figés.

Le premier critère est un critère référentiel qui suppose qu'à une unité lexicale donnée correspond un référent unique. Ce critère peut prendre appui sur cette citation de M. Grevisse (*1964)* dans « *Le Bon Usage* »:

« Un mot, quoique formé d'éléments graphiquement indépendants, est composé dés le moment où il évoque dans l'esprit, non les images distinctes répondant à chacun des mots composants, mais une image unique. Ainsi les composés ''hôtel de ville'', ''pomme de terre'', ''arc de triomphe'', éveillent chacun dans l'esprit une image unique et non les images distinctes d'''hôtel'' et de ''ville'', de ''pomme'' et de ''terre'', et d'''arc'' et de ''triomphe''. »

Le second critère proposé est d'ordre sémantique. Ce critère adopte ce postulat : le sens du composé n'est pas compositionnel. Mais il s'avère inefficace dans la mesure où il existe des exemples de composés où le sens est compositionnel. Dans une expression telle que *« faire chou blanc »,* les éléments composants ne conservent pas leurs sens : il n'y a donc pas compositionnalité du sens. En revanche dans un syntagme nominal tel que *« chaise longue »* ou *« mauvaise herbe »*, nous retrouvons bien le sens des éléments composants qui contribuent au sens général du syntagme et c'est l'ajout d'autres sens que ceux des composants qui confère au syntagme le statut de composé, autrement dit de syntagme figé.

Le troisième critère est d'ordre syntaxique : une séquence figée implique que les opérations syntaxiques normalement disponibles dans les séquences libres soient bloquées pour les séquences figées. Par exemple, il se peut que des modifications syntagmatiques comme l'ajout d'un modifieur soient impossibles.

Ex 1.4.2.a) : « une chaise longue ».

*Ex 1.4.2.b) : « * une chaise | inexorablement | longue ».*

*Ex 1.4.2.c) : « * une chaise | très | longue ».*

Dans les exemples *Ex 1.4.2.b* et *Ex 1.4.2.c*, il n'est pas possible d'interpréter ces groupes nominaux comme des syntagmes figés.

Même s'ils permettent de se faire une idée de ce qui caractérise principalement le figement, ces critères restent tout de même génériques et devraient être plus précis.

- *Critères proposés par Gaston Gross*

Pour Gaston Gross, une séquence figée est une suite de mots ayant une existence autonome. Cette définition permet d'opposer le figement à la dérivation.

Une séquence figée peut donc offrir deux lectures. Prenons l'exemple suivant :

Ex 1.4.2.d) : « Les carottes sont cuites ».

Cette phrase peut avoir pour interprétation :

Ex 1.4.2.d)(i): « Les carottes sont cuites »

⇒ Les légumes sont prêts.

Dans cette interprétation, le sens est compositionnel.

La seconde interprétation notée *(ii)* n'est pas prédictible à partir des éléments constituant la séquence.

Ex 1.4.2.d)(ii): « Les carottes sont cuites » :

⇒ La situation est désespérée.

Dans cette dernière phrase, le sens est donc non compositionnel.

Gaston Gross introduit la notion d'opacité sémantique pour décrire la phrase *(ii).* L'opacité sémantique est un des critères proposés par G. Gross pour distinguer les séquences libres des séquences figées. Ce critère correspond au critère de compositionnalité proposé au début de la section. Une suite est dite opaque quand le sens n'est pas compositionnel et à l'inverse cette suite est dite transparente quand le sens est compositionnel. Pour Gaston Gross, la phrase *(ii)* est donc opaque ou « sémantiquement figée et contrainte lexicalement ».

Les séquences figées permettent donc aux locuteurs d'avoir une double lecture d'un même énoncé. En effet un énoncé peut être interprété de manière compositionnelle ou figée. Le contexte d'énonciation permet aux locuteurs d'interpréter convenablement un énoncé : il n'y a donc pas d'ambiguïté entre ces deux lectures.

G. Gross propose un deuxième critère qui est le blocage des propriétés transformationnelles. Ce critère correspond au critère syntaxique proposé au début de la section. Les séquences libres tolèrent en général un certain nombre de transformations ou changements de structure. Ainsi des transformations telles que la passivation, la pronominalisation, le détachement, l'extraction ou la relativation qui sont des transformations courantes sont rendues impossibles.

Prenons par exemple l'expression « casser sa pipe ». Cette expression, dans sa lecture compositionnelle ou sémantiquement transparente, signifie « briser l'objet qui sert à fumer et qui est une pipe ». La lecture figée ou sémantiquement opaque de cette expression a pour sens « mourir ».

Nous allons donc confronter les propriétés transformationnelles de la construction libre et de la construction figée ; pour cela nous allons prendre appui sur les deux phrases suivantes :

1 : « *Pierre a cassé son stylo* » ⇨ cette construction est libre : le substantif « pipe » a été remplacé par le substantif « stylo » pour mieux mettre en évidence les différences entre construction libre et construction figée.

2 : « *Pierre a cassé sa pipe* » ⇨ cette construction est figée et revient à dire « Pierre est mort ».

Comme nous allons le voir, l'application des diverses transformations est possible sur la phrase 1 qui est une construction libre.

Phrase 1 : « Pierre a cassé son stylo »	
Transformations syntaxiques	Résultats obtenus
Passivation	? Son stylo a été cassé par Pierre (Le stylo de Pierre a été cassé)
Pronominalisation	Pierre l'a cassé
Détachement	Son stylo, Pierre l'a cassé
Extraction	C'est son stylo que Pierre a cassé
Relativation	Le stylo que Pierre a cassé

Tableau 1.4.2.a) : Tests transformationnels appliqués à la phrase « Pierre a cassé son stylo »

Les différentes phrases obtenues sont grammaticales et donc possibles. Le point d'interrogation « ? » qui précède la phrase 1 après la passivation indique que cette phrase est grammaticale mais est difficilement acceptable pour un locuteur natif et peu courante dans la langue ; néanmoins ce signe n'interdit pas cette phrase.

Ces mêmes transformations opérées sur la phrase figée sont impossibles ou ne sont pas naturelles:

Phrase 2 : « Pierre a cassé sa pipe » (« Pierre est mort »)	
Transformations syntaxiques	Résultats obtenus

Passivation	? Sa pipe a été cassée par Pierre (*La pipe de Pierre a été cassée)
Pronominalisation	*Pierre l'a cassée
Détachement	Sa pipe, Pierre l'a cassée
Extraction	*C'est sa pipe que Pierre a cassée
Relativation	*La pipe que Pierre a cassée

Tableau 1.4.2.b) : Tests transformationnels appliqués à la phrase « Pierre a cassé sa pipe »

Le signe « * » indique que ces phrases ne sont pas correctes.
Les structures obtenues après les transformations peuvent être considérées comme étant grammaticales, c'est-à-dire correspondant aux règles syntaxiques communes mais elles interdisent la lecture figée : seule la lecture compositionnelle est tolérée. Mais si l'on considère uniquement la lecture figée alors ces phrases sont impossibles. Seule l'opération syntaxique du détachement conserve la lecture figée de la phrase 2.

Il est toutefois possible de constater que certaines constructions n'admettent pas certaines transformations mais que ces constructions ne soient pas figées pour autant. G. Gross prend pour exemple le mot « regarder » dans le sens synonyme de « concerner » pour souligner cette idée. Les exemples présentés ci-dessous sont ceux proposés par G. Gross :

« concerner »	Actif	Cette affaire nous concerne tous
	Passif	Nous sommes tous concernés par cette affaire
« regarder »	Actif	Cette affaire nous regarde tous
	Passif	*Nous sommes tous regardés par cette affaire

Tableau 1.4.2.c) : Différence structurelle entre deux verbes synonymes : « concerner » et « regarder »

Gaston Gross met en parallèle ces deux constructions dont les verbes sont synonymes. Nous pouvons alors voir que la construction avec le verbe « regarder » n'admet pas la passivation mais il s'agit bien de constructions libres. G. Gross en vient donc à conclure et insiste sur le fait que « l'opacité sémantique et les restriction syntaxiques vont de pair ». Ces critères ne doivent donc pas être considérés séparément.

Un dernier critère prolonge l'opacité sémantique et le blocage des propriétés transformationnelles, il s'agit de la non-actualisation des éléments. Ce critère pose le principe

suivant : une suite est composée « quand aucun des éléments lexicaux constitutifs ne peut être actualisé ».
L'actualisation permet d'inscrire un prédicat dans son contexte. Un verbe par exemple est actualisé par sa conjugaison.

L'actualisation des éléments est permise dans les séquences libres mais non dans les séquences figées. L'exemple qui suit illustre ce critère :

Lecture compositionnelle :

Ex 1.4.2.e):

Pierre a pris une veste ⇨ *Pierre a pris un vêtement*

Pierre a pris sa veste

Pierre a pris cette veste

Lecture figée:

Ex 1.4.2.e):

Pierre a pris une veste ⇨ *Pierre a été battu aux élections*

**Pierre a pris sa veste*

**Pierre a pris cette veste*

L'actualisation du mot « veste », qui se fait au moyen de la détermination, n'est pas possible pour la séquence figée.

Gross utilise le terme de « locution » pour désigner une suite de mots dont les éléments constitutifs ne sont pas actualisés. Nous reviendrons plus longuement sur la définition de ce terme.

L'impossibilité de substituer un mot d'une séquence par un autre mot appartenant à la même classe sémantique ou par un synonyme permet aussi de distinguer les séquences figées des séquences libres. Dans l'expression « casser sa pipe », il n'est pas possible de substituer le verbe « casser » par le verbe « briser ».

De même, les séquences figées n'acceptent pas en général l'insertion d'éléments nouveaux. Les modifieurs sont souvent interdits. Par exemple, l'expression « *tourner de l'œil* » n'admet pas des séquences telles que :

*« *il tourne de l'œil gauche »,*

*« *il tourne d'un seul œil ».*

Mais il est cependant possible d'insérer un modifieur après le terme qui porte les marques de la flexion :

« Il tourne vraiment de l'œil ».

Le tableau suivant énumère les différents critères proposés par G. Gross qui caractérise les expressions figées.

Critères principaux	
1	Opacité sémantique
2	Blocage des propriétés transformationnelles
3	Non-actualisation des éléments constitutifs de l'expression
4	Substitution synonymique impossible
5	Non-insertion d'éléments nouveaux

Tableau 1.4.2.d) : Principaux critères de reconnaissance des expressions figées

Ces différentes caractéristiques du figement que nous venons de passer en revue ne sont pourtant pas communes à toutes les expressions figées. Il s'agit là de quelques propriétés générales qui constituent des indices pour la reconnaissance d'expressions figées. Mais les critères permettant de reconnaître des locutions adjectivales figées ne seront pas les mêmes que ceux employés pour reconnaître des locutions verbales figées.

Une autre notion intervient dans la description du figement : il s'agit de la notion de « degré de figement ». Salah Mejri, dans son ouvrage consacré au *« Figement lexical » (1997),* remarque que le figement s'inscrit dans un continuum : en effet le « passage des S.L. (Séquences Libres) s'opère d'une manière graduelle et imperceptible aux S.F. (Séquences Figées) ». Une séquence dite « figée » n'est jamais totalement figée ou a contrario entièrement libre. Nous retrouvons chez Gaston Gross cette notion de degré de figement. Les séquences à noyau verbal que nous allons étudier illustrent tout à fait cette notion : en effet le figement n'atteint pas les propriétés morphosyntaxiques du verbe. Le verbe connaît donc toutes les variations qui lui sont propres dans le cadre de ces séquences. Pour Maurice Gross, la différence entre les Séquences Figées et les Séquences Libres réside dans la « saturation lexicale de certaines positions ». En effet, les expressions figées ne sont jamais entièrement figées, seuls certains éléments de ces expressions sont contraints.

1.5 Les différents types d'expressions figées

D'après les différents travaux menés sur les expressions figées, il est possible de dénombrer six principaux types d'expressions figées : les noms composés, les déterminants composés, les locutions adjectivales, les locutions conjonctives et prépositives, et enfin les locutions verbales.

Les noms composés

Il s'agit là du type d'expression figée le plus courant dans la langue et le plus étudié par les linguistes. Sa dénomination le prouve : en effet le *nom composé* est le seul type de séquence figée à bénéficier d'un terme spécifique tandis que les différents autres types de séquences figées sont traditionnellement regroupés sous le terme générique de *locution*. La composition constitue avec la dérivation une des principaux moyens de formation des nouveaux mots comme nous l'avons vu précédemment. Le nom composé est donc un mot construit qui se range sous la catégorie « mot polylexical » du schéma (cf. *figure 1.4.1.a*) et s'oppose donc au mot dérivé. Le nom composé met en jeu des éléments lexicaux autonomes.

Il est communément admis que le trait d'union permet de reconnaître les noms composés ; c'est le cas pour des mots tels que « porte-monnaie » ou « porte-manteau » mais ce critère s'avère inefficace. La soudure est un des problèmes qui se pose pour la reconnaissance des noms composés : les éléments lexicaux constituants sont collés les uns aux autres. Le mot « surenchère » par exemple est un mot composé alors qu'au premier abord, on pourrait déduire qu'il s'agit d'un mot dérivé.

La composition implique l'opacité sémantique : des groupes nominaux dont le sens serait compositionnel ne peuvent donc pas être considérés comme des composés.

Les déterminants composés

Il est possible de distinguer deux types de détermination : une détermination simple et une détermination complexe. La détermination simple se fait au moyen d'articles définis ou indéfinis, d'adjectifs possessifs ou démonstratifs. La détermination complexe aussi dite polylexicale est discontinue et met en jeu plusieurs mots.

La catégorie des déterminants composés comprend aussi les modifieurs figés. Il s'agit en général de complément de nom exprimant l'intensité ou un trait qualificatif.

Ex 1.6.a) : elle a une peau $_{SP}$ [de bébé].

Dans cet exemple, le modifieur, c'est-à-dire le syntagme prépositionnel, est figé.

Les locutions adjectivales

Ce terme de locution désigne des adjectifs composés dits aussi « adjectivaux ». Comme nous l'avons vu dans la partie *1.4.1 (Figement et composition)*, le figement et la composition ne sont pas synonymes et ne désignent pas le même phénomène.

Dans l'exemple suivant, la locution adjectivale est constituée par un adjectif composé :

Ex 1.6.b) : Ce travail est à faire.

Au premier abord, il peut sembler étrange que le groupe de mots « à faire » forme une locution adjectivale dans la mesure où la tendance générale parlerait plutôt d'un syntagme

prépositionnel. Mais le fait que cette suite soit pronominalisable par le pronom « le » plutôt que par le pronom « en » indique que la suite est de nature adjectivale.

Ex 1.6.c) : Ce travail est à faire et le sera encore demain.

Dans cette construction adjectivale, le sens est compositionnel : il est possible de prédire le sens de la suite à partir des éléments lexicaux qui la constituent. Il ne s'agit donc pas d'une construction figée.
Dans une séquence comme « au parfum », le sens n'est pas compositionnel : en effet rien ne prédit que cette suite signifie « être au courant ». Il s'agit donc là d'un adjectif composé figé.

Les locutions adverbiales

Les adverbes simples sont à distinguer des adverbes complexes ou polylexicaux, c'est-à-dire constitués de plusieurs éléments lexicaux. Ces adverbes complexes, quand ils ont un fonctionnement régulier, peuvent être reformulés par des paraphrases :

Ex 1.6.d) : Il marche rapidement.

Ex 1.6.e) : Il marche avec rapidité.

Les suites adverbiales figées connaissent les restrictions communes aux expressions figées : la substitution synonymique est impossible ou limitée, le éléments sont sémantiquement opaques. C'est le cas dans les exemples suivants :

(Marcher) à reculons.

(Boire) à tire-larigot.

Les locutions prépositives et conjonctives

Les prépositions introduisent des compléments d'un verbe transitif indirect (Complément d'objet indirect) ou d'un verbe à deux compléments (Complément d'objet second).
Les conjonctions introduisent des propositions complétives.

Ces deux parties du discours ont donc un fonctionnement parallèle. Des expressions telles que « au fur et à mesure que » (locution conjonctive) et « à l'instar de » (locution prépositive) sont considérées comme étant figées.

Les locutions verbales que nous allons étudier plus en détail dans la section suivante sont le dernier type d'expressions figées qu'il nous reste à présenter.

2. Les verbes et les Locutions Verbales

2.1 Les verbes

Maurice Gross (*1988*), dans un article intitulé « Les limites de la phrase figée » met à jour une tripartition des verbes. Cette tripartition correspond en fait aux différentes natures sémantiques de la fonction verbale. Les trois types de verbes qu'il est donc possible de distinguer sont : les verbes usuels, les verbes composés, et les verbes supports.

Les verbes usuels

Les verbes usuels regroupent des verbes classiques tels que « manger » ou « donner ». Les verbes sont des prédicats qui peuvent être actualisés par leur conjugaison et par leurs compléments. Le verbe « manger » suppose que le sujet du verbe soit un humain et le complément du verbe soit un nom appartenant à la classe sémantique de la nourriture. Cette relation combinatoire entre le verbe et son complément peut être représentée ainsi :

MANGER (h, n)
h = humain
n = nourriture

Le verbe opère une sélection sur l'ensemble des noms et cette sélection est restreinte dans la mesure où n'importe nom ne peut pas se combiner avec n'importe quel verbe. Ce type de notation sera utilisé dorénavant dans ce travail pour décrire un prédicat et ses arguments.

Les verbes composés

M. Gross utilise le terme de « verbes composés » pour désigner les verbes qui apparaissent dans des expressions figées. Cet adjectif « composé » permet de signifier que les expressions dans lesquelles figure ce type de verbes sont non compositionnelles du point de vue sémantique, c'est-à-dire que le sens de ces expressions n'est pas prédictible. Il est donc aussi possible de parler de verbe figé. Une propriété soulignée par de nombreux auteurs réside dans la possibilité pour un verbe figé d'être substitué par un verbe ordinaire. En effet, le verbe qui apparaît dans une phrase figée, et le ou les compléments avec lesquels il est employé peut être substitué par un verbe morphologiquement simple et sémantiquement équivalent.

Ex 2.1.a) : Max casse du sucre sur le dos de Luc

≈ Max dénigre Luc

Le verbe composé et ses compléments « casser du sucre sur le dos de » peut être substituer par le verbe simple « dénigrer ».

Les verbes supports

Il s'agit de verbes qui, comme leur nom l'indique, servent de support à des prédicats nominaux. Les verbes supports sont des verbes de sens général qui n'ont pas de fonction prédicative, et qui apportent à un substantif prédicatif les informations de temps, de personne, et de nombre et des informations aspectuelles. Le verbe support et le prédicat nominal avec lequel il est construit peuvent également être paraphrasés par un verbe simple sémantiquement équivalent.

Ex 2.1.b): Faire un voyage ➪ Voyager
Donner l'autorisation ➪ Autoriser

D'après Gaston Gross, le verbe support permet d'actualiser le prédicat « tout comme le fait la désinence verbale avec le prédicat verbal ». G. Gross, par voie de conséquence parle aussi de noms supports :

Ex 2.1.c): Douanier ➪ Agent des douanes

Des verbes comme le verbe « être », « avoir », ou « faire » ont généralement un emploi de verbe support. D'autres verbes peuvent être connotés sémantiquement et apportent aux substantifs une actualisation mais aussi une contribution sémantique.

Ex 2.1.d): Max déborde d'affection pour Marie.

Le verbe « déborder » semble avoir un emploi de verbe support. Maurice Gross ajoute d'ailleurs que les nominalisations sont des transformations qui transforment des phrases à verbes ordinaires en phrases à verbes supports.

Ex 2.1.e): Max juge sévèrement Luc.
Max porte un jugement sévère sur Luc.

Les verbes supports ne présentent pas de restriction de sélection sur l'ensemble des noms comme c'est le cas des verbes ordinaires. Gaston Gross dans un article consacré à la lexicographie (1981) utilise le terme de « verbes opérateurs » pour faire allusion aux verbes supports. Il dit donc qu'un verbe simple est traduit par un verbe de sens général désigné sous l'appellation de « verbe opérateur » accompagné par un substantif de même racine que le verbe simple.

Ex 2.1.f): Le juge a lu le verdict.
Le juge a donné lecture du verdict.

Chacun des trois types de verbes décrits ici présentent des caractéristiques particulières. Il est cependant impossible de proposer des classes de verbes parfaitement distinctes. En effet, ces trois classes de verbes sont très proches et peuvent se confondre dans la mesure où les trois types de constructions ont recours au même lexique.

Ex 2.1.g.1): Max porte une caisse.

Ex 2.1.g.2): Max ne porte pas Luc dans son cœur.

Ex 2.1.g.3): Max porte de l'affection à Luc.

Le verbe « porter » est présent dans ces trois phrases mais il n'a cependant pas le même emploi. Dans la phrase *g.1)*, il s'agit d'un verbe ordinaire, la phrase n'est donc pas figée et le verbe « porter » a un emploi libre. Dans la phrase *g.2)*, en revanche le verbe « porter » est un verbe composé employé dans une phrase figée et la phrase *g.3)* emploie ce verbe en tant que support du nom « affection ». Comme nous pouvons le voir, un même verbe présente les trois emplois décrits ci-dessus. Le contexte peut permettre de déterminer s'il s'agit d'un verbe ordinaire, d'un verbe support, ou d'un verbe figé.

Maurice Gross souligne cependant que certaines expressions figées ont recours à un vocabulaire spécifique, c'est-à-dire à des éléments lexicaux spécifiques. Le nom « escampette », par exemple, apparaît uniquement dans l'expression « prendre la poudre d'escampette » qui équivaut au verbe simple « fuir ».

Il est donc particulièrement difficile de distinguer clairement ces trois catégories de verbes qui présentent de nombreux points communs ; c'est la raison pour laquelle certains auteurs ne font pas de distinction particulière entre ces verbes. En effet, certains auteurs, comme Hervé Curat ne prennent en considération que deux types de verbes : les verbes ordinaires et les verbes composés qui incluent aussi bien les verbes figés que les verbes supports.

2 .2 La notion de locution

La locution s'apparente à une « formule déjà construite, préfabriquée », d'après Blanche-Noëlle Grunig (*1997*). « *Le Dictionnaire de Linguistique et des Sciences du Langage* » (*1994*) donne la définition suivante :

Locution : « la locution est un groupe de mots (nominal, verbal, adjectival) dont la syntaxe particulière donne à ces groupes le caractère d'expression figée et qui correspondent à des mots uniques. Ainsi « faire grâce » est une locution verbale (ou verbe composé), correspondant à « gracier » […] ; « mise en jeu » est une locution nominale (ou nom composé). »

Alain Rey (*1977*) donne une définition de la locution dans une perspective lexicographique :

Locution : « la locution est unité fonctionnelle plus longue que le mot graphique et appartenant au code de la langue (devant être apprise) en tant que forme stable et soumise aux règles syntactiques […] L'expression est cette même réalité considérée comme « une manière d'exprimer quelque chose » ; elle implique une rhétorique et une stylistique. »

Alain Rey exclut du champ de la phraséologie les dictons et les proverbes (locutions-phrases), les mots complexes (locutions fonctionnelles).

Les notions de « locution » et de « mot composé » sont des notions vagues et difficiles à définir avec précision. Elles sont aussi souvent amalgamées. La définition du *« Dictionnaire de linguistique et des Sciences du Langage »* semble supposer que les deux termes sont synonymes et qu'il est possible d'employer l'un ou l'autre indifféremment. Pour notre part, nous considérerons que le terme de locution ne s'applique pas pour les constructions nominales figées que nous appellerons « noms composés » et non « locution nominale ».

Etymologiquement le mot « locution » signifie « manière de dire ». La tradition grammaticale attribue l'appellation de « locution » à des séquences inférieures au niveau de la phrase. En effet, les phrases entièrement figées sont généralement appelées expressions idiomatiques. On parle donc de locutions verbales, adjectivales, ou prépositionnelles…

David Gaatone (*1991*) estime que « si elle se présente quantitativement comme une séquence de mots, la locution apparaît intuitivement comme l'équivalent d'un mot unique ». Puis il ajoute que le fait d'attribuer ce terme spécifique de « locution » permet justement d'infirmer que ce groupe de mots n'est pas assimilable à un mot unique même si la possibilité de lui trouver un « équivalent plus ou moins approximatif sous forme de mot unique, et qu'on doive, en outre la considérer comme unité lexicale ». Certains auteurs utilisent le terme de « coalescence » pour souligner cette impression d'équivalence entre une locution verbale et un verbe simple. Hervé Curat (*1986*) ajoute à ce sujet que la locution verbale laisse une impression de verbes en deux mots. En effet, cette intuition d'équivalence entre un élément d'une locution et un élément simple aurait pour origine la cohésion particulièrement forte qui existe entre les mots composants de la locution, cohésion que l'on ne retrouve pas dans un « syntagme ordinaire ». De plus les éléments qui apparaissent dans une locution donnée peuvent apparaître dans d'autres contextes et avoir un emploi libre, c'est la raison pour laquelle les auteurs ont recours au terme de « locution » pour désigner ces groupes de mots qui présentent une très forte cohésion et dont les éléments constituants ne sont pas dissociables.

Les termes de « mot composé » ou « locution » renvoient généralement à des notions spécifiques chez certains auteurs. Gaston Gross parle par exemple d'unités polylexicales pour désigner des unités morphologiquement complexes. Danielle Corbin (*1997*), quant à elle, parle d'unités polylexématiques pour désigner ces unités lexicales complexes dans la mesure où elle estime que les termes « unités polylexicales » et « lexie » sont inappropriés. Ces unités polylexématiques que décrit D. Corbin selon deux propriétés principales sont traditionnellement rangées sous l'étiquette « locution » ou « mot composé ». L'une des propriétés est d'ordre syntaxique et réside dans le fait que l'unité polylexématique peut occuper dans la phrase une position de constituant syntaxiquement minimal autonome. Il est à noter que D. Corbin exclut les collocations de par cette propriété. Nous verrons dans le paragraphe suivant quelle est la définition de la collocation et ce qu'elle représente par rapport aux expressions figées.

Robert Martin, pour sa part, définit simplement la locution comme étant un « syntagme figé situé au-delà du mot et en-deça de la phrase figée. » Les locutions sont traditionnellement opposées aux syntagmes libres. Thun (*1978*) dit en effet que « la locution est l'aboutissement, dans une synchronie donnée, d'un processus de figement, de pétrification, de fossilisation ». Le caractère non-compositionnel du sens dans ces locutions a

longtemps été l'un des principaux critères qui permettrait d'identifier une séquence figée. Nous étudierons dans la section suivante quels sont les différents critères proposés par les auteurs pour reconnaître les locutions verbales.

2.3 Traits caractéristiques des locutions verbales figées

Il est possible de voir au travers des différents travaux produits par les différents auteurs que le terme de « locution verbale » ne désigne pas nécessairement les mêmes séquences selon leurs théories respectives. Certains auteurs regroupent sous ce terme aussi bien des séquences figées à noyau verbal que des constructions à verbe support. Lars Lindberg (1898) définit quant à lui une locution verbale comme « une proposition où le verbe s'est affaibli ou a perdu son caractère de verbe, et où tous les mots se sont rapprochés pour former ensemble une unité ». Il ajoute que ces locutions sont de sens général et qu'elles peuvent se figer graduellement ou subitement. La diachronie prend donc une place importante dans son analyse. Il est à noter que d'après sa définition de la locution verbale figée comme étant une « locution contenant un verbe à un mode personnel, et qui a pris une forme fixe, et perdu dans une certaine mesure son caractère primitif » désigne aussi des séquences telles que « voici » ou « voilà » qui sont des formes ayant qui résultent d'un processus de soudure. Tenter de fournir une définition précise et formelle se révèle donc être une tâche difficile. Georges Bernard dans un article publié en 1974 mesurait déjà l'ampleur de cette difficulté : « c'est presque une gageure que de prétendre définir les caractéristiques formelles des locutions verbales. »

Gaston Gross propose deux principaux critères pour reconnaître une locution verbale : le sens de la suite de mots doit être non compositionnel, c'est-à-dire opaque, et syntaxiquement contrainte, les modifications et transformations doivent donc être impossibles. Ces deux propriétés étaient déjà données par G. Gross pour distinguer les séquences libres des séquences figées. Mais dans sa description des locutions verbales, il est possible de trouver davantage de précisions.

En effet, un des principaux problèmes qui se pose dans la définition d'une locution verbale réside dans les divergences théoriques des auteurs. Nous avons effectivement vu dans la section consacrée aux verbes que les verbes supports bien que très ressemblants étaient à distinguer des verbes figés, cependant certains auteurs regroupent sous ce même terme de « locution verbale » aussi bien des séquences figées que des constructions à verbe support. Des auteurs comme Hervé Curat ou David Gaatone ne font pas la distinction entre ces deux types de constructions. Guilbert qui utilise le terme « d'unité syntaxique verbale » pour désigner des séquences verbales figées décrit le verbe comme étant une sorte « d'opérateur servant à la transformation d'un nom en verbe ». Les unités syntaxiques verbales d'après cette définition désigneraient donc des constructions à verbe support et non des séquences verbales figées. Il réserve en effet le terme de « locution verbale » à des « unités phraséologiques », des manières de parler incrustées dans le lexique de la langue par l'usage constant ». D'après G. Gross, le fait de ne pas discriminer ces deux types de constructions a d'importantes implications théoriques. Il étudie plus particulièrement deux phrases dont les structures paraissent identiques :

1. *Pierre a faim.*

2. Pierre a froid.

Dans ces deux phrases, le sujet est un être humain animé et le nom qui suit le verbe est abstrait. Le verbe employé dans ces constructions est le verbe « avoir » qui est souvent utilisé en tant que support permettant d'actualiser prédicat nominal. A première vue, nous pourrions donc supposer qu'il s'agit de deux constructions à verbe support. Mais quelques tests vont nous permettre de constater que ces deux constructions sont différentes. Ces tests et les résultats produits sur chacune des deux phrases sont représentés dans le tableau *2.3.a)* :

	Pierre a faim	**Pierre a froid**
Insertion de modifieurs	Pierre a très faim	Pierre a très froid
	Pierre a plus faim qu'hier	Pierre a plus froid qu'hier
	Pierre a une faim de loup	* Pierre a un froid de canard
Relativation	La faim que Pierre a	* Le froid que Pierre a
Génitif	La faim de Pierre	* Le froid de Pierre

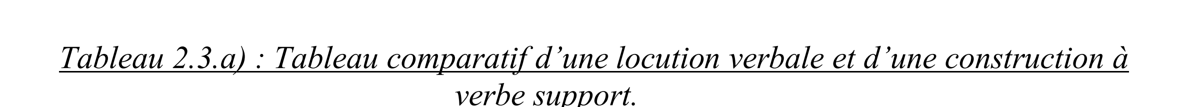


Tableau 2.3.a) : Tableau comparatif d'une locution verbale et d'une construction à verbe support.

Le test de la relativation est déterminant pour savoir si une suite est figée ou s'il s'agit d'une construction à verbe support. G. Gross ajoute que la relativation n'est généralement pas possible quand la détermination est figée. La phrase « Pierre a pris la fuite » est justement un exemple de phrase où le déterminant est figé. Après relativation cette phrase devient « *la fuite que Pierre a prise ». La phrase obtenue est agrammaticale parce que la détermination est figée comme le montrent les tests ci-dessous :

Prendre la fuite.

** Prendre| une| fuite.*

** Prendre| sa | fuite.*

** Prendre|des |fuites.*

Il semblerait donc qu'il ne s'agisse pas d'une construction à verbe support mais d'une locution verbale figée. Cela peut sembler étrange dans la mesure où le sens de la suite est

compositionnel et non opaque, de plus le verbe est de sens général, mais la détermination étant contrainte, la suite est considérée comme étant figée.

La compositionnalité, ou plus précisément la non-compositionnalité, a, chez la plupart des auteurs, constitué le critère principal de reconnaissance des expressions figées comme nous l'avons déjà dit dans la section précédente. Robert Martin (*1994*) définit la non-compositionnalité comme un phénomène observable essentiellement en synchronie dans la mesure où « historiquement la non-compositionnalité n'existe pas ». L'opacité sémantique proviendrait de démotivations étymologiques. L'expression « porter le chapeau » aurait par exemple pour origine une coutume médiévale qui consistait à faire porter un chapeau ridicule aux personnes condamnées et de les promener ensuite à travers les cités. Gaatone, par exemple, parle de « non calculabilité du sens » : « le sens global d'une locution en général et d'une locution verbale en particulier ne peut être obtenu par l'addition du sens des constituants individuels, comme cela se ferait pour un syntagme verbal ». Chez Gaatone, le terme « syntagme » désigne une suite libre alors que le terme « locution verbale » renvoie à une suite figée. Comme nous l'avons vu dans les sections précédentes, une même suite peut offrir deux lectures. C'est ce que nous pouvons remarquer dans un exemple proposé et analysé par Roy (*1976*) :

« Mettre au pied du mur »

1ère lecture : ≈ *« déposer au pied du mur »*

⇨ *Syntagme verbal*

2ème lecture : ≈ *« forcer à prendre parti »*

⇨ *Locution*

Dans la première lecture, le sens est obtenu par la « combinaison des éléments lexicaux et d'éléments grammaticaux et prosodiques ». Tandis que la deuxième interprétation résulte d'un emploi métaphorique.

Pour Gaatone, « la locution verbale constitue sur le plan sémantique un tout inanalysable ». Ce « tout inanalysable » parait supposer qu'aucune règle ne régit cet ensemble. Georges Bernard précise cependant que l'ensemble constitué par les locutions verbales n'est aucunement un « ensemble anarchique et aléatoire ». En effet, les locutions verbales présentent une structure interne et ne sont donc pas entièrement figées. Les constructions ne présentant pas de structure interne sont à un stade de figement. D'après Gaatone, ces constructions entreraient dans un état proche de la soudure entre les mots.

Pendant longtemps et surtout dans les anciens travaux, l'absence de déterminant constituait le critère essentiel dans la reconnaissance des locutions verbales. C'est le cas de Hervé Curat qui pense donc que des séquences du type « avoir peur » ou « prendre fin », … qui sont en fait des constructions à verbe support seraient des locutions verbales dans la mesure où le verbe est directement suivi par son complément qui n'est pas modifié par un déterminant. L'insertion de modifieurs n'est permise que si le déterminant n'est pas figé. Une grande majorité des expressions figées sont non compositionnelles du point de vue sémantique, les déterminants employés dans ces séquences n'ont donc aucune contribution sémantique. L'absence d'article présage que le substantif qui apparaît dans la séquence figée ne fait pas référence et ne renvoie à aucun objet de la réalité. Gougenheim (*1971*) fait partie

de ces auteurs qui estiment qu'une séquence donnée est figée dés lors que le substantif qui apparaît dans cette séquence n'est pas introduit par un déterminant : il considère en effet que l'absence de déterminant est le seul critère nécessaire pour définir une locution verbale. Guilbert (*1961*) nuance cette hypothèse en précisant que l'absence de déterminant ne constitue jamais qu'un indice de figement et ne permet pas de définir une locution verbale. Ce critère est aussi celui choisi par Kayne (*1975*) ou Ruwet (*1975*). Maurice Gross (*1985*) précise toutefois que l'absence de déterminant constitue seulement une marque de figement en se basant sur des données statistiques. En effet, le nombre de locutions sans déterminant serait assez réduit et se limiterait à 1800 constructions sur les 8000 séquences verbales figées dénombrées. De plus, certaines constructions libres tolèrent l'absence de déterminant. Il ne s'agit donc pas d'une caractéristique propre aux locutions verbales.

Georges Bernard est un des premiers auteurs à proposer des critères de classement des locutions verbales. L'utilisation de deux principaux critères permet d'aboutir à l'émergence de quatre classes distinctes. La possibilité ou non pour une construction donnée d'avoir une expansion, et la possibilité ou non pour cette construction de commuter avec une construction articulée constituent ces critères de classement. Ces différentes classes et les exemples donnés par Georges Bernard sont représentés dans le tableau ci-dessous *(2.3.b)* :

	Classes de Locution Verbales	*Exemples proposés par G. Bernard*
1	Locutions non commutables et saturées	*Porter plainte, plier bagage, lâcher prise, …*
2	Locutions non commutables et non saturées	*Donner lieu à, faire main basse sur, …*
3	Locutions commutables et saturées	*Faire l'appel, …*
4	Locutions commutables et non saturées	*Faire l'affaire de, …*

Tableau 2.3.b) : Classement des locutions verbales proposé par Georges Bernard.

Le terme saturation est employé par Georges Bernard pour indiquer si la construction est à même de recevoir une expansion ou non. Cette notion de saturation pourrait être assimilée à celle de transitivité. En effet, les constructions saturées sont constituées du verbe et d'un complément direct tandis que les constructions non saturées se composent d'un verbe, de son complément et d'une préposition pour introduire un second complément. La notion de commutatibilité ou d'articulation renvoie au fait que le complément du verbe de la locution est modifié par un article. Les locutions non commutables et saturées qui ne permettent

aucune modification ou insertion pourraient être considérées comme un modèle de figement maximum d'après Georges Bernard.

Thun (*1978*) propose aussi de classer les séquences verbales selon qu'elles soient sémantiquement transparentes ou sémantiquement opaques. Parmi les séquences sémantiquement opaques, il est possible de retrouver les locutions qui s'opposent aux syntagmes libres.

Gaston Gross (*1996*), propose une définition précise et caractéristique de la locution verbale dans son ouvrage consacré aux expressions figées. Le terme de locution verbale désignerait donc toute suite composée d'un verbe et de ses compléments qui présenterait une non compositionnalité du sens ou le figement des groupes nominaux, c'est-à-dire que les groupes nominaux compléments ne peuvent subir aucune modification. La substitution des déterminants et l'insertion de modifieurs adverbiaux sont donc impossibles. La locution verbale s'oppose à un syntagme verbal libre dont la seule contrainte sera la sélection du domaine d'argument par le verbe. Les phrases entièrement figées et les constructions à verbe support ne sont pas considérées comme étant des locutions verbales par Gaston Gross.

Les locutions verbales présentent de nombreuses ressemblances avec les syntagmes verbaux libres. Gaatone réfute précisément pour cette raison l'utilisation du terme de « locution verbale » et préfère utiliser le terme de « locution » pour désigner des séquences figées dont l'élément principal n'est pas un verbe. Le verbe aussi bien dans une locution verbale que dans un syntagme peut se construire avec un complément d'objet direct ou un complément d'objet indirect. Les locutions verbales ont donc la même structure interne que les syntagmes libres. Les structures spécifiques que pourraient présenter les locutions verbales sont rares voir inexistantes. Le verbe qui apparaît dans une locution verbale connaît toutes les modifications morphologiques qui lui sont propres, comme c'est le cas dans un syntagme verbal libre. La forme verbale peut donc être fléchie normalement qu'il s'agisse d'un verbe libre ou d'un verbe figé. Le degré de figement des locutions verbales n'est pas le même pour toutes les locutions verbales, il y a donc un continuum entre les syntagmes libres et les locutions verbales figées.

Malgré ces similitudes, il existe tout de même quelques différences entre les locutions verbales et les syntagmes libres. En effet, dans une construction libre, les éléments qui apparaissent en position d'argument peuvent être des classes d'objets. D'après Maurice Gross (*1998*) représentent des « classes sémantiques construites à partir de critères syntaxiques ». Un exemple de classe d'objet est la « nourriture » qui constitue le domaine d'argument du verbe « manger ». Un prédicat donné sélectionne donc son argument dans une classe d'objet donnée. Pour les locutions verbales, en revanche, les positions d'argument sont remplies par des éléments isolés et non par des classes d'objet. De même, contrairement aux syntagmes libres, l'actualisation des compléments n'est pas possible. L'actualisation concerne principalement la détermination, comme nous l'avons vu. Dans la plupart des locutions verbales, il est possible d'observer que la détermination est figée.

Le critère suivant qui permet de différencier les locutions verbales des syntagmes verbaux est un critère formel. Le critère formel est généralement celui que les auteurs privilégient dans leurs théories. En effet, des séries de tests qui diffèrent selon les auteurs sont proposées pour reconnaître les locutions verbales. Les tests de transformation proposés ici

sont ceux décrits dans les travaux de Gaston Gross repris des travaux de Maurice Gross. Ces tests nous paraissent plus pertinents que ceux proposés par d'autres auteurs dans la mesure où G. Gross distingue les locutions verbales des constructions à verbe support. G. Gross propose donc six tests transformationnels pour déterminer si une suite est syntaxiquement contrainte.

Le passif est le test formel le plus communément cité par les différents auteurs bien que ceux-ci reconnaissent qu'il s'agit d'un test d'une efficacité relative et qu'il est loin d'être suffisant pour reconnaître une locution verbale.

Ex 2.3.a): *Pierre a pris la tangente.*
⇨ *La tangente a été prise par Pierre.*

Une suite figée ne peut pas être transformée par la passivation. Dans le cas des suites qui proposent une double lecture, seule la suite proposant une lecture compositionnelle admettra le passif.

Ex 2.3.b): *prendre la mouche.*

Lecture compositionnelle : attraper

Pierre a pris la mouche ⇨ *la mouche a été prise par Pierre.*

Lecture figée : se vexer

Pierre a pris la mouche ⇨ **la mouche a été prise par Pierre.*

Nous pourrions donc supposer que toute suite qui ne peut pas être transformée par la passivation est figée. Mais les suites figées ne sont pas les seules constructions qui ne peuvent pas être passivées. Les constructions employant des verbes intransitifs ne tolèrent pas le passif. Ce test ne permet donc pas de reconnaître de manière catégorique une séquence figée.

L'extraction est le second test proposé par G. Gross. Cette transformation consiste dans un changement de structure qui s'applique à un argument (sujet ou objet) pour opposer deux éléments dans un paradigme donné.

Ex 2.3.c):

Séquence non figée *: prendre la route.*

Pierre a pris la route ⇨ *C'est la route que Pierre a prise.*

Séquence figée *: prendre la tangente.*

Pierre a pris la tangente ⇨ **C'est la tangente que Pierre a prise.*

Comme nous pouvons le voir dans ces exemples, la deuxième phrase comportant une locution verbale ne permet pas l'extraction du complément. En effet, le fait que les positions argumentales dans une locution verbale ne soient pas remplies par des classes paradigmatiques rend l'extraction impossible.

Un autre test qui permet de reconnaître les locutions verbales est le détachement qui permet de mettre en évidence un élément dans une suite donnée. Le détachement est donc une transformation très proche de la focalisation. Il est possible de remarquer que le détachement

n'est possible que si le déterminant est défini : ceci est valable aussi bien pour les suites figées que les suites libres.

Ex 2.3.d):
Séquence non figée : *acheter un manteau.*
*Pierre a acheté un manteau ⇨ *Un manteau, Pierre l'a acheté.*

Séquence figée : *prendre la tangente.*
*Pierre a pris la tangente ⇨ *la tangente, Pierre l'a prise.*

Les locutions verbales ne permettent pas cette transformation : comme le montre la phrase ci-dessus, le détachement est donc impossible pour les suites verbales figées.

La pronominalisation (ou reprise anaphorique) n'est pas permise pour les locutions verbales dans la mesure où les pronoms qui sont employés dans ces séquences n'observent pas le même fonctionnement que les vrais pronoms.

Ex 2.3.e):
Séquence non figée : *acheter un manteau.*
Pierre a acheté ce manteau dans un magasin chic.
⇨ Ce manteau, Pierre l'a acheté dans un magasin chic.

Séquence figée : *prendre le large.*
Pierre a pris le large.
*⇨ *Luc l'a pris (le large).*

Les groupes nominaux dans les locutions verbales ne renvoyant pas précisément à des objets concrets, la pronominalisation est impossible.

La relativation que nous avons vu dans un paragraphe précédent est fort utile pour différencier les locutions verbales des constructions à verbe support. La relativation n'est pas possible pour les locutions verbales.

Ex 2.3.f):
Séquence non figée : *prendre une décision.*
Pierre a pris une décision ⇨ La décision que Pierre a prise.

Séquence figée : *prendre la tangente.*
*Pierre a pris la tangente ⇨ *La tangente que Pierre a prise.*

Il est aussi possible de procéder au test de l'interrogation pour reconnaître les locutions verbales. En effet, cette transformation ne peut s'appliquer que sur des structures libres.

Ex 2.3.g) :
Séquence non figée : *prendre un livre.*
Pierre a pris un livre
⇨ *Qu'est-ce que Pierre a pris ? Un livre.*

Séquence figée : *prendre la tangente.*
Pierre a pris la tangente
⇨ *Qu'est ce que Pierre a pris ? *La tangente.*

Le tableau ci-dessous récapitule d'une manière plus schématique les différents tests proposés par G. Gross :

Transformation		*Locution Verbale*	*Séquences libres*
1	*Passif*	-	+
2	*Extraction*	-	+
3	*Détachement*	-	+
4	*Pronominalisation*	-	+
5	*Relativation*	-	+
6	*Interrogation*	-	+

Tableau 2.3.c) : Critères proposés par Gaston Gross pour la reconnaissance des locutions verbales.

Les suites verbales pour lesquelles aucune transformation n'est possible auront donc un sens opaque. La possibilité ou non d'opérer ces transformations permet d'échelonner le degré de figement des locutions verbales : en effet, une suite est d'autant plus figée que le nombre de transformations syntaxiques possibles est réduit.

Au-delà de ces différents tests, il est possible d'observer quelques autres traits formels caractéristiques des locutions verbales. Le fait que le nombre soit invariable dans les locutions constitue l'un de ces traits. En effet, une grande majorité des expressions figées emploient des groupes nominaux au singulier. Le passage au pluriel rend généralement ces séquences agrammaticales ou leur fait perdre leur interprétation figée.

Ex 2.3.h) : prendre la mouche.
Lecture compositionnelle : attraper
Pierre a pris la mouche ⇨ *Pierre a pris les mouches.*

Lecture figée : se vexer
Pierre a pris la mouche ⇨ **Pierre a pris les mouches.*

Des auteurs comme Rohrer (*1967*) ou Björkman (*1978*) ont proposé dans leurs travaux respectifs des listes de tests formels très précis pour déterminer si une séquence donnée constitue un bloc figé ne permettant qu'un nombre limité de transformations.

Il peut malgré ces différents indices être difficile de reconnaître les locutions verbales car elles ne se résument pas à un verbe figé employé avec un complément quelconque. En effet, c'est la combinaison formée par le verbe et son complément introduit ou non par un déterminant et figée à des degrés divers qui va constituer une locution verbale.

3. La terminologie adoptée

Comme nous l'avons vu tout au long de cette partie, les termes employés pour décrire le figement et les locutions sont nombreuses et variées et la terminologie employée par les différents auteurs souligne des points de vue théoriques différents.

Dans le cadre de ce travail, une expression figée sera désignée par le terme de « séquence ». Nous parlerons donc plus précisément de séquences verbales figées. Danielle Corbin a cependant souligné dans ses travaux que le terme de « séquence » supposait une continuité entre les formes, or il est possible d'observer une absence de continuité entre les formes simples (mais morphologiquement complexes) et les formes composées de plusieurs mots, cela est d'autant plus vrai pour les locutions verbales. Ce terme de « séquence », également employé par S. Mejri, présente toutefois l'avantage d'être un terme neutre qui montre bien que divers domaines d'études entrent en jeu pour le traitement du figement.

Nous parlerons donc également de « locutions verbales », malgré les réticences d'auteurs comme D. Gaatone ou A. Rey pour désigner les séquences verbales montrant un degré quelconque de figement. L'adjectif « figé » ne sera donc que très peu utilisé avec le terme de « locution verbale » pour éviter toute redondance.

La notion de « mot » sera aussi utilisée mais dans sa définition la plus basique à savoir dans sa définition typographique. Un mot représentera donc une unité enclavée par deux blancs (ou espaces).

Comme nous l'avons vu dans un paragraphe précédent, les termes de « collocation » ou « cooccurrence » sont souvent utilisés pour désigner des séquences figées. Mais ces termes supposent un degré moindre de figement. Hausmann (*1985*) définit la collocation comme une « combinaison polaire non arbitraire de deux lexèmes qui a un caractère conventionnel à l'intérieur d'un groupe linguistique ». La collocation se présente donc bel et bien comme une séquence figée mais elle représente le degré le plus faible de figement dans la mesure où elle respecte la compositionnalité du sens et elle ne fait que contraindre la liberté de cooccurrence. Une séquence comme « créer un fichier » est un exemple de collocation. Dans une perspective de Traitement Automatique, une collocation pourrait se définir comme des séquences de mots anormalement récurrentes et qui correspondent à des associations statistiques préférentielles. Ce terme n'est donc pas tout approprié pour désigner les séquences verbales figées. En effet, le terme de « locution verbale » présente l'avantage (ou l'inconvénient selon les points de vue) de ne pas donner d'indication sur le degré de figement de la séquence qu'il représente et est donc générique.

Un terme, dans le domaine linguistique de la terminologie, représente un signe linguistique qui accompagne l'apparition d'un nouveau concept dans un domaine donné. Le mot « terme » dans cette première partie n'a pas été utilisé dans son acception la plus neutre et

commune à savoir l'équivalent du « mot ». Nous verrons que dans la seconde partie, la notion de « terme » sera aussi bien utilisée dans son acception linguistique à savoir en tant qu'objet d'étude de la terminologie que dans son acception classique.

Nous verrons que le figement et les locutions verbales posent des problèmes particuliers pour le traitement automatique. Nous allons voir dans la seconde partie plus en détail la construction de l'application Verbalex à proprement parler.

Partie B

Les expressions figées et les locutions verbales, du point de vue du Traitement Automatique des Langues

I/ TAL: méthodologies et outils pour l'analyse automatique des expressions figées est des locutions verbales

Nous avons vu précédemment les traits linguistiques du figement et des locutions verbales. Nous allons étudier dans cette partie les méthodes et deux différents outils qui proposent un traitement des expressions figées dans une perspective d'analyse automatique. Nous verrons également d'une manière plus détaillée les différentes étapes de construction du logiciel Verbalex.

1. Le traitement informatique des Séquences Figées

Si la littérature est plutôt abondante et hésitante en ce qui concerne le traitement linguistique des expressions figées, les choses sont tout autres dans le domaine du traitement automatique. Nous verrons donc dans un premier temps que les travaux du LADL ont fortement contribué à donner aux expressions figées l'importance qui leur était due. Nous étudierons ensuite quelques méthodologies possibles pour reconnaître automatiquement les expressions figées.

1.1 Les travaux du LADL

Le LADL (Laboratoire d'Automatique Documentaire et de Linguistique) a été fondé en *1967* par Maurice Gross. Ce centre d'études s'est ensuite consacré à l'étude des expressions figées dans une perspective de traitement automatique. Les travaux du LADL se proposent de fournir de manière systématique une description des expressions figées, aussi bien d'un point de vue syntaxique que sémantique. Ces travaux s'inscrivent dans la lignée de ceux de Z.S. Harris qui portaient sur la théorie transformationnelle. En effet, Z.S. Harris (*1988*) considérait les « phrases élémentaires ou noyaux comme unités de base de la composition syntaxique ». Les phrases élémentaires seraient sémantiquement invariantes par transformation. Cette hypothèse issue des travaux de Harris a été intégrée à la théorie du lexique-grammaire initiée par le LADL. Cette théorie consiste dans l'étude systématique pour tous les mots du lexique d'une langue donnée de leurs propriétés syntaxiques. Cela reviendrait plus précisément à étudier dans quelles constructions syntaxiques entre chaque mot, d'où l'appellation « lexique-grammaire ». Cette théorie écarte l'approche sémantique jugée trop subjective et variante d'un linguiste à l'autre.

L'intégration de l'hypothèse transformationnelle à la théorie du lexique-grammaire suppose l'émergence de deux questions devant aboutir à deux faits précis. Il faut dans un premier temps en arriver à conclure que tout mot entre dans une phrase élémentaire caractéristique et qu'il n'a donc aucune autonomie syntactico-sémantique. Dans un second temps, la conclusion qui s'impose consiste dans le fait que toute phrase complexe s'analyse en terme de phrases élémentaires.

C'est dans ces idées que résidaient les principales préoccupations du LADL du moins à son fondement. Ces dernières ont ensuite évoluées pour finalement s'éloigner totalement des fondements du cadre génératif transformationnel « instauré » par Chomsky et d'autres auteurs. La raison de cet éloignement réside principalement dans le fait que les perspectives génératives transformationnelles étudient des phénomènes syntaxiques indépendamment d'une étude du lexique pour aboutir à une formalisation et des généralisations, ces généralisations seraient basées sur un nombre d'exemples réduit ou plus précisément insuffisant.

Le LADL a alors orienté ses travaux dans une démarche de description du fonctionnement concret des mots du lexique. Le LADL s'est donc consacré à une étude « extensive » et « intensive » du lexique du français. Cette étude est dite extensive dans la mesure où elle a pour objet la majeure partie du lexique et intensive car elle prend à cœur de mettre jour le maximum de propriétés connues pour chaque item lexical qui compose le lexique.

Cette étude du lexique se base sur des phrases élémentaires construites pour être analysées. Les énoncés déjà existants qu'ils soient oraux ou écrits ne sont pas utilisés car ils sont généralement trop longs et sources d'ambiguïtés multiples. Les phrases sont tout d'abord soumises à un jugement d'acceptabilité pour déterminer si la phrase élémentaire construite est grammaticale ou ne l'est pas. Les mots qui composent ces phrases sont ensuite analysés selon leur contexte et leurs cooccurrences. L'étude d'un mot donné aboutit à l'émergence d'un certain nombre de propriétés. Des professionnels de la linguistique ont ensuite pour tâche de valider les propriétés définies pour chaque mot.

Les travaux du LADL ont ensuite naturellement évolué vers une étude extensive des expressions figées du français, naturellement car une étude de chaque mot du lexique ne pouvait que laisser présager une telle démarche. En effet, les expressions figées sont constituées de mots simples qui ont donc selon les contextes tantôt un emploi libre tantôt un emploi figé. Il était donc « logique » que les expressions figées prennent davantage d'importance dans l'approche entamée par le LADL. Le LADL a donc entrepris de recenser toutes les expressions figées, ce qui a permis de mesurer au sens propre le poids de ces expressions dont le nombre est nettement supérieur à celui des formes libres. L'ensemble des études menées par le LADL aura donc eu pour conséquence de faire du figement un objet linguistique autonome.

Les divers résultats produits par le LADL sont représentés sous la forme de tables. Ces différentes tables représentent le lexique-grammaire élaboré au LADL. Les tables qui composent ce lexique-grammaire regroupent tous les éléments du lexique. Chacune de ces tables contient un ensemble de propriétés qui s'établissent en colonne. En vis-à-vis de ces colonnes, un codage avec un signe positif « + » ou négatif « - » permet de préciser si

l'élément du lexique figurant dans la table peut être défini ou non par cette(s) propriété(s). Ces tables constituent actuellement des fichiers excel au format « .xls » pour constituer des ressources électroniques.

Nous pouvons donc constater que le LADL a une démarche morphologique quant au traitement des corpus. En effet, le filtrage des séquences complexes correspond à la reconnaissance des propriétés syntaxiques définies par les tables.

Silberztein (*1987*) précise donc que « la reconnaissance purement lexicale n'est plus de mise » particulièrement dans le cas des locutions verbales car le verbe connaît de grandes variations qu'elles soient flexionnelles ou transformationnelles.

1.2 Méthodologies

La reconnaissance automatique des expressions figées pose des problèmes tout à fait spécifiques. Les dernières décennies ont assisté à la naissance et à l'évolution de diverses méthodes qui ont plus ou moins fait leurs preuves. Nous allons donc voir dans cette section les diverses approches possibles pour la reconnaissance automatique des séquences figées.

1.2.1 La « Zone Fixe » des expressions figées

Le traitement automatique d'un corpus présuppose que tous les mots qui figurent dans ce corpus soient connus et identifiés. Une opération d'étiquetage a donc une grande importance dans le cadre d'une analyse automatique. Cette opération consiste donc à fournir des informations morphosyntaxiques sur les mots qui composent les différentes phrases d'un corpus donné. Les informations fournies par l'étiquetage différeront selon les buts visés par l'analyse. Un même mot peut aussi bien avoir un emploi libre qu'un emploi figé : l'étiquetage ne prend en compte que les mots simples. Cette étape à elle seule ne permet donc pas de reconnaître les locutions verbales.

Eric Laporte (*1988*) introduit la notion de zone fixe pour décrire un mode de reconnaissance automatique des expressions figées. La « zone fixe » d'une expression figée désignerait la partie de l'expression qui admet un nombre de fixe mots simples, même si ces mots sont susceptibles de variations morphologiques ». Dans le cas des séquences verbales, les verbes supports sont exclus de la zone fixe. Dans une expression telle que « être bon public », la zone fixe se limitera à « bon public » dans la mesure où le verbe « être » est un verbe support et qu'il peut donc être effacé ou remplacé par une variante aspectuelle.

Ex 1.2.1.a) : être bon public : Pierre est bon public

- *Pierre a des amis bons publics. (effacement)*
- *Pierre est devenu bon public. (remplacement par une variante aspectuelle)*

⇨ *ETRE = Verbe Support.*

La reconnaissance de la zone fixe permettrait d'aboutir à la constitution d'une base de données contenant les formes de différentes expressions figées existantes ainsi que leurs propriétés. Cette méthodologie permettrait de reconnaître automatiquement les expressions

figées dans la mesure où les formes données par le dictionnaire apportent aussi des informations distributionnelles. La zone fixe d'une locution comme « casser sa pipe » serait donc décrite de la manière suivante :

`N0 Casser Poss pipe`

N0 est une variable désignant le groupe nominal Sujet. Le possessif *Poss* est donc variable.

La zone fixe permet donc de reconnaître des séquences figées même lorsque ces dernières connaissent des variations.

Mais reconnaître cette zone fixe ne revient pas à assurer la présence dans le corpus traité de la locution correspondante. Dans certains cas, repérer la zone fixe d'une expression figée peut s'avérer suffisant pour affirmer que cette expression figure dans le corpus : c'est le cas quand un mot n'apparaît que dans le cadre de la locution et qu'il n'a pas d'emploi libre, comme c'est le cas dans l'exemple suivant :

Ex 1.2.1.b) : N0 prendre la poudre d'escampette.

Le mot « escampette » n'existe pas en dehors de cette locution.

Mais ces exemples de locutions ne sont pas les plus répandus et la simple reconnaissance de la zone fixe s'avère donc insuffisante. En effet, certaines contraintes formelles pèsent sur des structures de phrase afin de garantir une lecture figée de la phrase en question. Nous reprendrons pour illustrer ce point l'exemple « casser sa pipe ». Un locuteur ne privilégiera l'interprétation figée que si l'adjectif possessif est coréférent au sujet libre N0 et si ces derniers s'accordent en genre et en nombre.

De plus, le contexte dans lequel apparaît l'expression joue un rôle d'importance dans la délimitation de la zone fixe de cette expression. Mais cela n'est en aucun cas déterminant pour savoir si l'expression figée figure dans le corpus. En effet, quand la zone fixe d'une expression figée est reconnue, l'hypothèse la plus plausible est que cette expression soit présente dans le corpus traité. Mais il ne s'agit pas là d'une certitude : une analyse est donc nécessaire pour confirmer ou infirmer la présence d'une locution. Eric Laporte conclut en précisant que « la reconnaissance de la zone fixe d'une expression figée apporte la présomption que celle-ci figure dans le texte ». Mais cette présomption ne peut être assimilée à une « information certaine ».

1.2.2 Les méthodes statistiques et / ou structurelles

Le traitement automatique met en avant l'utilisation de deux principales méthodes dans l'analyse des expressions figées dans un corpus donné : une approche statistique ou une approche structurelle. La méthode statistique a la particularité de ne nécessiter qu'un nombre limité de connaissances linguistiques. L'approche structurelle, quant à elle, requiert davantage de connaissances linguistiques. En effet, un outil statistique n'a recours qu'à un lexique de mots fléchis et leurs catégories pour assigner des étiquettes grammaticales aux mots d'un texte. L'outil structurel a au minimum besoin de grammaires locales de la langue du texte. Nous fournirons plus avant une définition aussi précise que possible des grammaires locales.

Les travaux récents ont tendance à allier ces deux méthodes statistique et structurelle pour produire des résultats les plus efficaces possibles. Ces nouvelles approches auront par exemple recours à un analyseur statistique puis à un étiqueteur structurel.

Les méthodes statistiques sont fréquemment utilisées dans le domaine du TAL dans de nombreux secteurs particuliers. Les résultats les plus satisfaisants se retrouvent notamment en acquisition lexicale par la recherche d'associations récurrentes entre mots voisins, ou encore en extraction d'information par l'attribution de mots-clés à des textes, ou en génération automatique de textes. Les analyses statistiques sont généralement réalisées sur des corpus quantitativement importants. Les résultats produits ne sont pas directement accessibles et appréciables. Des fonctions mathématiques spécifiques permettent de déceler dans les corpus des associations préférentielles. Les méthodes statistiques présentent l'avantage de mettre au même plan différents niveaux d'analyse. Elles présentent cependant un inconvénient de taille : les données émanant des résultats produits sont difficiles à exploiter.

Contrairement aux méthodes statistiques, les méthodes structurelles nécessitent des connaissances linguistiques les plus précises et les plus complètes possibles avant de procéder au traitement du corpus. La rencontre d'un mot inconnu dans le corpus devient donc problématique et fait échouer le traitement : d'autres niveaux d'analyse sont alors nécessaires pour remédier à ces lacunes. Les opérations de filtrages après ce traitement sont sensiblement réduites. Les méthodes statistiques si elles ne font pas appel à des connaissances linguistiques avant le traitement nécessite des opérations de filtrage plus importantes après ce traitement pour produire des résultats plus probants.

De nombreux projets d'outils d'extraction d'information terminologique ont recours à l'une ou à l'autre de ces méthodes ou à des méthodes hybrides.

1.2.3 L'acquisition de termes en terminologie : présentation de quelques outils

De nombreuses applications ont pour objectif l'extraction d'information d'ordre terminologique. Ces applications sont connues dans le domaine du TAL et font généralement référence. Des outils tels que ACABIT (*Daille, 1994*), LEXTER (*Bourigault, 1994*) et DicAssist permettent l'acquisition de termes. Nous allons donc voir quel type de méthode utilisent ces outils et quel est leur mode de fonctionnement.

1.2.3.1 DicAssist

DicAssist est un système visant la construction et l'accès à une base de données d'expressions figées à partir des ressources de la Toile. Ce système s'appuie sur les ressources de l'Internet pour s'adapter à l'évolution constante et rapide du langage. Les différents corpus passés en traitement dans les outils terminologiques ne tiennent pas compte de ce dynamisme permanent et font preuve de statisme.

L'architecture du système DicAssist permet de contrôler toutes les étapes de traitement nécessaires à la constitution et à la gestion d'une base de données de ce type.

DicAssist présente donc une architecture dite modulaire dans la mesure où elle fait appel à différents serveurs, bases de données et programmes qui créent une unique chaîne de traitement. Un programme temporisé dénommé Webget permet de récupérer tous les nouveaux documents édités par un portail donné de la Toile. Le Webget a été configuré pour extraire tous les nouveaux articles d'un quotidien en langue portugaise. Ces textes sont enregistrés et répertoriés dans une base de données. Chacun des textes est alors traité individuellement afin d'extraire un ensemble de termes candidats. Un serveur dénommé SENTA permet de procéder à ce processus d'extraction. La liste de termes candidats est produite par SENTA est enregistrée dans une base de données comportant les expressions figées potentielles qui seront ensuite validées manuellement et enrichies linguistiquement. La validation des expressions tient compte d'informations contextuelles et statistiques. La validation d'une expression entraîne l'enrichissement linguistique de celle-ci. Cette phase consiste à associer chaque expression aux différents types d'expressions figées proposés par Gaston Gross (*1996*), c'est-à-dire catégoriser ces expressions en tant que noms composés, locutions verbales, locutions adjectivales, déterminants composés, locutions adverbiales ou locutions prépositives ou conjonctives. Des informations morphosyntaxiques sont apportées grâce au dictionnaire électronique POLLUX de langue portugaise afin de constituer une base de données d'expressions figées très complète.

Le logiciel SENTA chargé d'extraire les termes candidats joue donc un rôle d'importance dans cette chaîne de traitement. SENTA acronyme de « Software for the Extraction of N-ary Textual Associations » est un logiciel visant l'extraction terminologique et qui a recours à une méthode probabiliste pour ce faire. Trois concepts essentiels participent au bon fonctionnement de ce logiciel : les modèles N-grams positionnels, la mesure d'association Expectative Mutuelle et un algorithme d'extraction GenLocalMaxs.

L'extraction est tout d'abord basée sur la construction de modèles N-grams positionnels. Un N-gram positionnel est une séquence ordonnée de N unités lexicales correspondant à une séquence d'un énoncé délimité par la taille d'un environnement. Le modèle de N-gram positionnel du logiciel SENTA a été calibré de manière à constituer un environnement de sept unités lexicales. Le calibrage a pour conséquence de ne construire que les N-grams positionnels tels que N=1…7. Le second concept, la mesure d'association Expectative Mutuelle permet de mesurer si les séquences établies par les modèles de N-grams positionnels construits à partir du texte constituent des expressions figées. En effet, un nouveau modèle probabiliste a été conçu afin de procéder au traitement statistique de séquences constituées de plus de deux unités lexicales. Ce nouveau modèle intitulé Expectative Mutuelle permet de mesurer le degré de cohésion qui lie entre eux les éléments constitutifs d'un N-gram positionnel.

L'Expectative Mutuelle est basée sur une notion essentielle, l'Expectative Normalisée. Cette dernière notion permet de mesurer à quel point la présence d'un mot est essentielle pour garantir une interprétation figée dans un N-gram positionnel. Plus un N-gram positionnel correspondant à une séquence du texte est figé, moins la perte d'un élément constituant sera tolérée : la valeur de l'Expectative Normalisée sera alors élevée. Une série de calculs de probabilités permettent d'obtenir cette valeur. Le calcul de l'Expectative Mutuelle ne tient pas cependant pas compte de la fréquence d'occurrence des séquences extraites, mais permet de mesurer le fort degré de cohésion qui peut lier les éléments constituants d'une expression figée.

L'algorithme de sélection GenLocalMaxs, troisième concept important du logiciel SENTA permet de ne retenir que les N-grams positionnels les plus pertinents et les plus aptes à constituer des séquences figées. Cet algorithme permet de sélectionner tout N-gram positionnel dont le degré d'association est un maximum local. Aucun seuil n'est donc prédéfini, seul les calculs faits précédemment ainsi qu'un nouveau calcul de probabilité permettent de consigner une séquence dans la liste de candidats termes.

DicAssist est donc une interface qui met en place une chaîne de traitements à partir de textes en ligne pour constituer une base de données d'expressions figées la plus actualisée possible.

1.2.3.2 ACABIT

D'autres outils visent l'extraction d'un type de termes correspondant à une catégorie syntaxique précise. Béatrice Daille (*2001*) s'est ainsi intéressée à l'identification d'adjectifs relationnels dans des corpus. Un programme d'extraction de terminologie nommé ACABIT (Daille, *1996*) est utilisé pour procéder à cette recherche. Ce programme procède tout comme DicAssist à l'extraction de termes candidats. Un score statistique est ensuite utilisé pour classer ces termes candidats qui correspondent à un profil particulier dans la mesure où le traitement concerne les adjectifs relationnels. Les séquences recherchées et extraites se trouvent donc limitées à deux unités lexicales pleines qui correspondent au Nom suivi d'un Adjectif. Ces séquences peuvent être plus longues quand le nom est également modifié par un groupe prépositionnel. Des patrons syntaxiques qui sont donc prédéfinis à partir d'une analyse linguistique des propriétés caractéristiques des adjectifs relationnels favorisent l'extraction des termes candidats la plus correcte et la plus efficace. Le corpus ouvert en entrée doit, avant tout traitement, être étiqueté et lemmatisé. Après cette étape de prétraitement du corpus, le programme a ensuite recours à des grammaires locales à base d'expressions régulières pour procéder à l'extraction. Les informations morphosyntaxiques associées à chaque mot et fournies par l'étiquetage sont donc essentielles pour procéder à l'extraction. La lemmatisation permet de définir des structures de base dénuées de toute variation pour chaque candidat. Nous apporterons davantage de précisions sur ces opérations d'étiquetage et de lemmatisation dans la section abordant la construction du programme Verbalex.

1.2.3.3 LEXTER

Didier Bourigault a développé en *1994* un logiciel visant l'acquisition et l'interprétation de terminologie. Cet outil nommé LEXTER a été construit à la Direction des Etudes et Recherches d'EDF afin de répondre à des besoins industriels bien précis. Ce logiciel prend donc en entrée un corpus de langue française. D. Bourigault considère que les méthodes d'extraction basées sur des critères de fréquence ne sont pas les plus appropriées et les plus efficaces pour l'extraction de termes complexes. En effet, Bourigault estime que des méthodes à caractère davantage linguistique seraient plus adaptées dans la mesure où elles font appel aux caractéristiques linguistiques et formelles du terme, ce qui permet d'obtenir les résultats les plus précis possibles. La méthode d'extraction de terminologie utilisée par LEXTER se fonde donc sur une analyse syntaxique qui répond à divers principes. Des calculs statistiques sont ensuite appliqués aux résultats pour davantage de précision. Ces calculs

permettent donc d'effectuer un filtrage sur la liste des candidats termes pour ne retenir que les termes complexes les plus pertinents.

Les trois outils que nous venons de présenter ont donc pour objectif l'extraction de terminologie ou de motifs précis à partir d'un corpus donné. Ces trois outils qui visent plus ou moins le même but ont cependant recours à des méthodes différentes qui ne s'inscrivent pas tout à fait dans le cadre strict des méthodologies soit statistiques soit structurelles. Comme nous l'avons vu, l'extraction de candidats termes avec DicAssist se fait indépendamment de toute information linguistique, ces informations sont apportées en aval du traitement. A l'inverse, ACABIT et LEXTER fonde leur fonctionnement sur le traitement linguistique du corpus, à savoir l'étiquetage, la lemmatisation ou encore l'analyse syntaxique du corpus pour procéder ensuite à des traitements statistiques pour affiner les résultats de l'extraction.

Il n'existe donc pas une méthode unique pour la reconnaissance automatique des séquences complexes.

La méthode de reconnaissance consistant dans le repérage de la « zone fixe » d'une expression figée proposée par Eric Laporte s'appuie sur les dictionnaires électroniques regroupant les diverses tables constituées par le lexique-grammaire élaboré par le LADL. Nous allons donc voir dans la section suivante quelles sont les caractéristiques des dictionnaires électroniques et ce qui les distinguent des dictionnaires traditionnels.

2. Les dictionnaires électroniques

Nous allons donc voir dans cette partie plus en détail les dictionnaires électroniques. En premier lieu, nous étudierons ce qui distingue la lexicographie traditionnelle des dictionnaires électroniques. Nous étudierons ensuite les principales propriétés du Dictionnaire Explicatif et Combinatoire (DEC) établi par Igor Mel'cuk et Alain Polguère (*1995)* qui présente l'avantage de tenir davantage compte des phénomènes phraséologiques qu'un dictionnaire classique.

2.1 Lexicographie vs Dictionnaires électroniques

La tradition divise généralement le dictionnaire et la grammaire qui toujours dans la tradition sont des outils normatifs indispensables pour décrire la langue et le bon usage. Le dictionnaire qui a pour objectif la description du lexique permet de recenser toutes les irrégularités d'une langue tandis que la grammaire établirait des règles décrivant la régularité et la stabilité de cette même langue. Cette vision des choses, cela va sans dire, est archaïque, idéalisée et donc très éloignée de la réalité. Le lexique n'est donc pas destiné à consigner uniquement les irrégularités et les idiosyncrasies de la langue. Malgré cette séparation quelque peu radicale du lexique et de la syntaxe, il est possible de constater que l'élaboration d'un dictionnaire fait appel à ces deux descriptions qui sont finalement davantage liées que totalement opposées.

L'élaboration de dictionnaires repose cependant fondamentalement sur la distinction entre le lexique et la syntaxe. D'après Gaston Gross, l'analyse syntaxique devrait précéder toute démarche lexicographique.

Le but premier d'un dictionnaire est la description des mots ou unités lexicales qui composent une langue. L'objectif serait plus précisément de donner le « sens » d'un mot. Le sens d'un énoncé constitué d'un certain nombre de mots résulterait donc de la somme des mots composant cet énoncé. Les dictionnaires ont généralement recours à la syntaxe pour illustrer les acceptions des mots décrits.

Une notion importante dans les dictionnaires concerne les catégories associées aux mots du dictionnaire. Gaston Gross dit par ailleurs que « tout dictionnaire repose sur la notion de catégorie ». La lexicographie n'exclut donc pas totalement les informations d'ordre syntaxique.

Cette question de la séparation lexique/syntaxe se pose particulièrement pour la description des séquences figées. Dans le cas des locutions verbales, la question peut effectivement se poser : une locution verbale doit-elle figurer dans un dictionnaire ou dans un ouvrage de grammaire ? En effet, ces séquences se trouvent à la limite de ces deux disciplines dans la mesure où elles sont régies par des règles syntaxiques régulièrement mais qu'elles sont sémantiquement équivalentes à une unique unité lexicale. Si une locution verbale est consignée dans un dictionnaire, sous quelle entrée doit-elle être décrite ? Est-ce sous l'entrée correspondant au verbe ou celle correspondant au nom qui constitue la tête du groupe nominal objet ? G. Gross (*1987*) estime que la solution adoptée par les dictionnaires n'est pas satisfaisante dans la mesure où les expressions figées figurent généralement en fin d'article pour un souligner un emploi figuré. G. Gross répond cependant à une de ces questions en indiquant que les expressions figées n'étant pas prédictibles doivent être décrites dans le dictionnaire : le sens de ces séquences ne pouvant être obtenu par celui des éléments constituants, ces expressions doivent faire l'objet d'une mémorisation comme lorsque l'on apprend un nouveau mot.

Un dictionnaire d'usage contient donc principalement des informations correspondant au lemme du mot, à sa catégorie syntaxique, à sa définition, à ses différentes acceptions et éventuellement quelques exemples.

Les dictionnaires électroniques sont généralement constitués d'une manière toute autre. En effet, les dictionnaires électroniques ne sont normalement pas destinés à être utilisés par un utilisateur humain comme l'affirme André Dugas (*1990*). Dans le cas des dictionnaires électroniques, l'utilisateur humain se substitue à un ordinateur-utilisateur. Ces dictionnaires, dans cette perspective, ne constituent donc « que » des ressources d'une utilité précieuse qui permettent l'exécution d'un programme informatique. Ces dictionnaires se présentent sous la forme de bases de données lexicales entièrement formalisées afin d'éviter toute ambiguïté lors d'un traitement automatique. Les propriétés lexicales définissant chaque entrée comportent les informations les plus précises et explicites possibles afin d'éviter l'échec de la reconnaissance automatique.

Les dictionnaires électroniques différent donc en de nombreux points des dictionnaires classiques. N'étant pas destinés à être utilisés par un être humain, ces dictionnaires doivent

donc être aussi complets que possible. Les informations données par ces dictionnaires doivent également être explicites : les dictionnaires classiques n'ont généralement pas la nécessité d'être explicites car ils font appel aux connaissances pragmatiques et à l'adaptabilité des utilisateurs.

Les informations fournies pas les dictionnaires électroniques sont totalement codées et exsangues d'information d'ordre sémantique, à savoir d'indication de sens. Ces informations sont en effet destinées à être exploitées par des programmes informatiques afin de procéder à des traitements dans les divers secteurs du TAL. Les caractéristiques que nous venons de définir s'appliquent particulièrement pour les travaux du LADL. Comme nous l'avons vu dans la section précédente, les résultats émanant des travaux du LADL sont représentés sous forme de tables qui représentent le lexique-grammaire, par ailleurs des dictionnaires électroniques ont aussi été élaborés par le TAL. Ces dictionnaires électroniques constituent des ressources sur lesquelles s'appuient des outils d'analyse ou d'acquisition de termes.

Les divers dictionnaires du LADL renvoient plus exactement aux dictionnaires DELA (Dictionnaire Electronique du LADL). La construction de ces dictionnaires est basée sur une définition purement formelle du mot simple, qui diffère totalement de la description morphologique que nous avons donnée dans la première partie (*1.4.1*). Un mot simple se réduit donc à « une unité de texte définie sur l'alphabet des codes ASCII et ne comportant aucun séparateur (ni trait d'union, ni blanc ni apostrophe). Cette définition tient donc fortement compte de la graphie des unités. L'alphabet des codes ASCII compte plus de vingt-six lettres dans la mesure où il comprend également les divers caractères accentués disponibles en français. Quarante et un caractères composent donc cet alphabet :

a b c d e f g h i j k l m n o p q r s t u v w x y z à â ç é è ë î ñ ô ö ù û ü

Max Silberztein (*1990*) précise que certaines lettres d'origine étrangère sont issues d'emprunts qui ont été intégrés au français.

Cette définition du mot simple aboutit à la définition suivante du mot composé : « un mot composé est une séquence de mots simples ». Les mots composés sont à distinguer des groupes libres de mots simples. Silberztein prend pour illustrer ce point les deux exemples suivants :

Ex : Cordon rouge et cordon bleu.

- *cordon rouge* ⇨ « cordon de couleur rouge »
  ↳ Groupe libre de mots simples.

- *cordon bleu* ⇨ « cordon de couleur bleue »
  ⇨ « bon cuisinier »
  ↳ Mot composé.

Le principe permettant de dissocier ces deux types de mots est le suivant :

Une séquence de mots simples est figée (ou composée) si l'une au moins de ses propriétés syntaxiques, distributionnelles ou sémantiques ne peut être déduite des propriétés de ses constituants.

De ces principes et définitions résulte une répartition des unités lexicales totalement différente de celle décrite dans le paragraphe *1.4.1 (Figement et Composition, figure 1.4.1.a)* comme nous le montre le schéma ci-dessous :

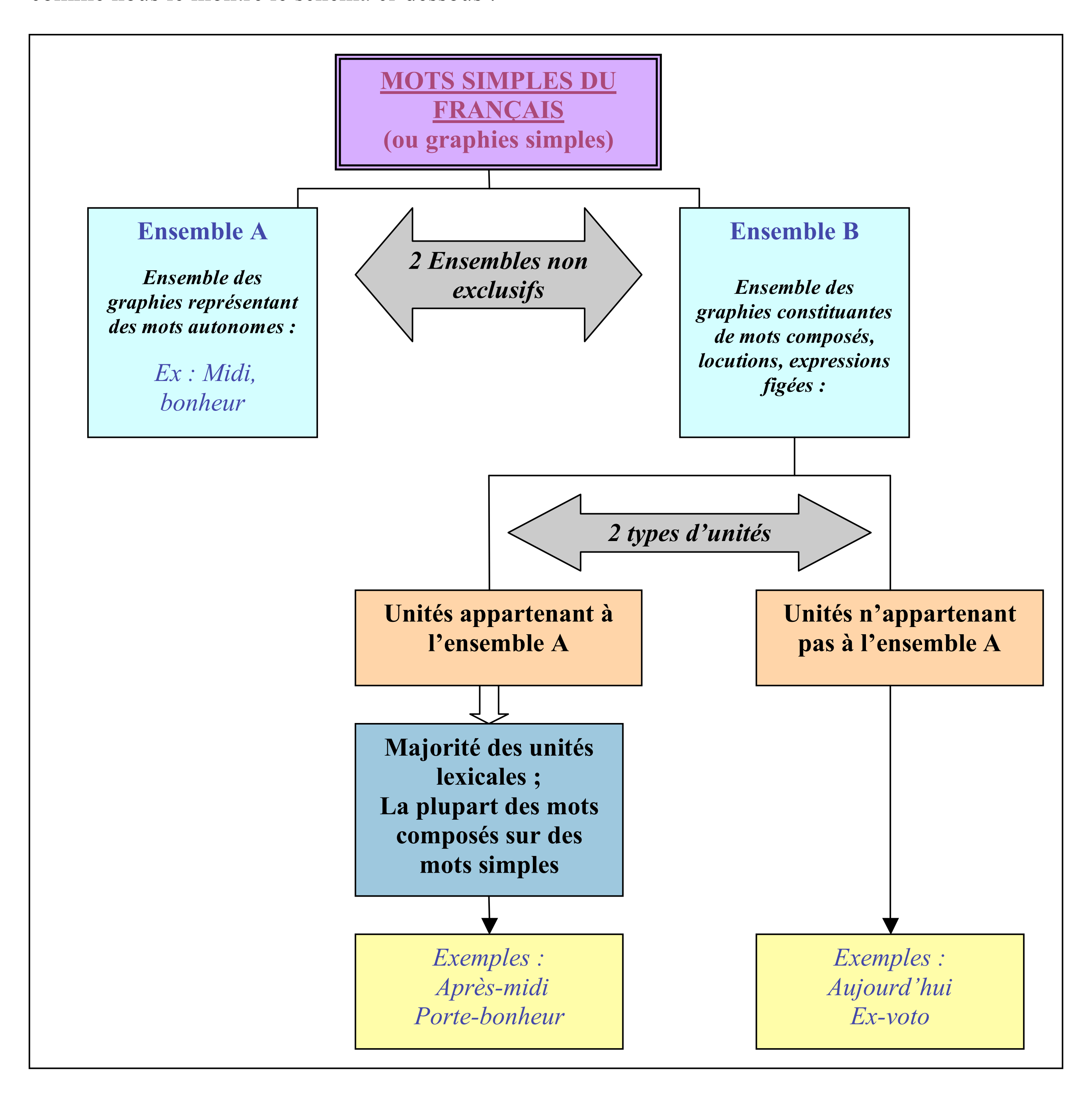


Figure 2.1.a): Les différents types d'unités lexicales dans la conception des dictionnaires électroniques du LADL

Les différents dictionnaires électroniques du LADL ou DELA sont au nombre de quatre. Le DELAS décrit donc la morphologie et la flexion des mots simples ; le DELAC a pour objet la description des mots composés ; le DELAF décrit les formes fléchies et les lexiques dérivés du français ; le DELACF est généré automatiquement à partir du DELAC pour décrire les formes fléchies et composées du lexique.

Ces différents dictionnaires électroniques constituent les ressources sur lesquelles s'appuient des logiciels tels que Intex ou Unitex dont nous décrirons le fonctionnement dans une prochaine section.

Ces dictionnaires que nous venons de décrire ne sont cependant pas les seuls disponibles. En effet, les versions numérisées des dictionnaires classiques initialement sur support papier reçoivent également l'appellation de dictionnaires électroniques. La version informatisée du TLF, autrement dit le Trésor de la langue Française, par exemple, est un dictionnaire électronique qui peut également constituer une ressource pour des programmes en TAL.

Les dictionnaires électroniques quels qu'ils soient se présentent généralement sous la forme de documents balisés ayant pour cela recours à des langages à structure balisante tels que XML ou SGML afin d'organiser et de hiérarchiser au mieux les informations selon leur pertinence.

Les dictionnaires issus de la pure tradition lexicographique aussi bien que les dictionnaires électroniques poursuivent donc le même objectif : décrire au mieux les unités lexicales d'une langue. Il s'agit aussi du but visé par le Dictionnaire Explicatif et Combinatoire (DEC).

2.2 Le Dictionnaire Explicatif et Combinatoire (DEC)

Le DEC ou Dictionnaire Explicatif et Combinatoire est un dictionnaire tout à fait différent et particulier par rapport aux dictionnaires classiques. Ce dictionnaire qui a pour auteur Igor Mel'cuk avec la collaboration d'Alain Polguère et André Clas tente de concilier approche logique et formelle dans l'étude des mots. Ce DEC résulte d'une démarche purement lexicologique et se distingue des dictionnaires issus de la lexicographie. Le DEC est le produit d'une théorie à part entière : il s'agit de la théorie Sens-Texte qui propose de partir d'une représentation sémantique pour construire des arbres syntaxiques à l'aide du lexique. La lexie est l'unité de base de cette étude lexicologique. Le terme « lexie » peut aussi bien désigner un mot simple qu'une locution. Ce mot ou cette locution sont pris en compte dans une acception spécifique. Si un mot donné est polysémique, le nombre de lexies disponibles pour ce même mot correspondra au nombre d'acceptions que ce mot reçoit. La lexie du DEC qui ne correspond pas tout à fait la « lexie » de Bernard Pottier comporte trois principaux composants : un sens, une forme graphique et phonique et un ensemble de traits combinatoires. Le dictionnaire est le produit final résultant de l'étude de l'ensemble des lexies d'une langue *L.*

Cette conception du lexique constitue une nouvelle approche et aura donc des conséquences importantes dans la conception des dictionnaires. Le DEFC (Dictionnaire Explicatif et Combinatoire du Français Contemporain) a donc été élaboré selon les principes établis par la théorie Sens-Texte. Cette théorie consiste à faire produire à un DEC toutes les informations qui pourraient permettre à un locuteur de « construire toutes les expressions linguistiques correctes de n'importe quelle pensée et ce, dans n'importe quel contexte ». Cette conception du DEC en fait essentiellement un dictionnaire de production.

Chaque lexie est décrite dans le DEC selon sa définition, ses connotations, et d'autres informations qui n'apparaissent pas ou alors apparaissent succinctement dans un dictionnaire classique. Un article du dictionnaire qui décrit la lexie doit comprendre dix zones principales apportant chacune une information sur la lexie, comme nous pouvons le voir dans le tableau ci-dessous :

ARTICLE : /LEXIE/		
1	**Zone vedette**	- Lexie vedette - Variante orthographique
2	**Zone phonologique**	- Prononciation - Prosodie particulière
3	**Zone morphologique**	- Partie du discours (ou catégorie) - Type de déclinaison ou de conjugaison - Formes irrégulières ou non réalisables
4	**Zone stylistique**	- Marques d'usage
5	**Zone sémantique**	- Définition - Connotations
6	**Zone de combinatoire syntaxique**	- Restrictions sur la cooccurrence syntaxique
7	**Zone de combinatoire lexicale restreinte**	- Restriction sur la cooccurrence lexicale
8	**Zone d'exemples**	- Exemples
9	**Zone phraséologique**	- Emplois figés
10	**Zone de Nota Bene**	- Remarques diverses

Figure 2.2.b): Les différents zones de description constituant un article du DEC

Ce dictionnaire décrit donc le lexique d'une langue sous la forme d'une énumération de lexies décrites selon les aspects définis que nous venons de définir. Une locution verbale

telle que « se mettre le doigt dans l'œil » constitue une entrée de dictionnaire au même titre que « se fourrer le doigt dans l'œil ».

Un projet d'informatisation du DEC est actuellement en cours afin de constituer un important dictionnaire électronique rassemblant les quatre volumes du DEC.

La structure du DEC permet donc d'intégrer d'une manière plutôt satisfaisante les locutions verbales et autres séquences figées. Nous verrons dans une prochaine section que le dictionnaire électronique produit par Verbalex est fortement inspiré par la structure du DEC.

3. Contraintes spécifiques liées aux locutions verbales

Le traitement des expressions figées et plus particulièrement celui des locutions verbales soulève des problématiques particulières au TAL. En effet, de nombreuses contraintes sont liées aux locutions verbales notamment en raison des variations que connaît le verbe dans le cadre de ces locutions.

Il est également possible de voir qu'un problème se pose pour la reconnaissance des locutions verbales : il s'agit de la discontinuité qui peut s'observer dans la locution. Certaines locutions tolèrent en effet l'insertion de modifieurs qui ne sont pas figés. Parmi ces modifieurs possibles figurent les propositions incises, des adverbes, des adjectifs ou encore des groupes prépositionnels. L'insertion d'une proposition incise qui est généralement enclavée de deux virgules constitue le cas le plus simple de discontinuité qui peut donc facilement être repérée automatiquement.

Ex 3.a): prendre en compte.
La police ne prend pas $_{ADV}$*[du tout] en compte ces preuves.*

Il est à remarquer que cet exemple est particulier : le groupe nominal objet peut être déplacé sans rendre la phrase agrammaticale :

Ex 3.b): prendre en compte.
La police ne prend pas ces preuves en compte.

Cette variante est acceptable mais relève peut-être davantage de l'oralité. Nous pouvons aussi nous demander à juste titre si la séquence « prendre en compte » constitue une locution verbale ou au contraire s'il s'agit d'une construction à verbe support.

	La police ne prend pas en compte ces preuves
Insertion de modifieurs	La police ne prend pas vraiment en compte ces preuves
	*La police ne prend pas en compte important ces preuves
	*La police ne prend pas, n'est-il pas vrai, en compte ces preuves
Relativation	*Le compte que prend Pierre de ces preuves
Génitif	*Le compte de Pierre de ces preuves



Locution Verbale

Figure 3.a): Tests de distinction entre locution verbale et construction à verbe support pour la séquence « prendre en compte »

Comme nous le montre la figure 3.a), l'insertion d'un adverbe ou d'une suite de type adverbiale est le seul type de modification toléré par la séquence « prendre en compte ». D'après ce que nous avons vu dans la première partie, les différentes transformations opérées sur cette séquence auraient été permises s'il s'était agi d'une construction à verbe support.

Parvenir à faire automatiquement cette distinction entre un verbe figé et un verbe support constitue l'une des contraintes liées au traitement des locutions verbales. En effet, un même verbe peut avoir tour à tour un emploi de verbe figé, de verbe libre ou bien de verbe support. Certains cas révèlent que seuls certains groupes nominaux compléments sont figés tandis que le verbe peut varier. Quelle étiquette peut-on attribuer à ce type de séquences ? Avons-nous affaire à une locution verbale ou à un nom composé ?

```
Ex 3.c): rater le coche.
Pierre a |raté    |le coche.
         | manqué |
         | loupé  |
```

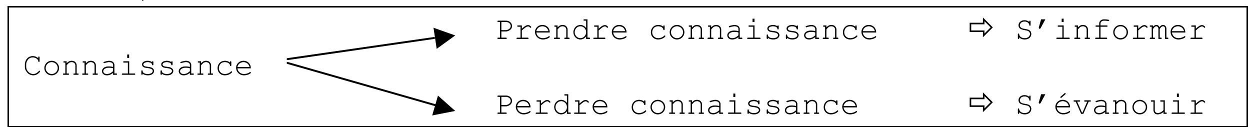
Dans cet exemple, nous pouvons donc voir que le groupe nominal objet est figé :

```
Ex 3.c): rater le coche.
Pierre a  raté *un coche.
Pierre a  raté *sa coche.
Pierre a  raté *le coche de sa vie.
```

Cet exemple comporte pourtant une locution verbale dans la mesure où même s'ils varient les verbes utilisés sont synonymes. C'est effectivement la combinaison du verbe qui a pour sens « rater » avec le groupe nominal objet « le coche » qui fait sens et qui permet d'accéder à

l'interprétation « rater une opportunité ». Un nom peut en effet totalement changer de sens selon le verbe avec lequel il est employé. Ainsi le nom « connaissance » produira un sens différent selon le verbe avec lequel il sera combiné.

Ex 3.d): connaissance.



La détermination est figée dans cet exemple proposé par Gougenheim (*1971*) : la présence de tout article est interdite.

De plus, une séquence ne doit pas être nécessairement entièrement figée pour être désignée par le terme de locution verbale. Il faut tenir compte des degrés de figement des locutions.

Une autre contrainte réside dans le choix des critères de reconnaissance : faut-il recourir à un critère purement formel (blocage des propriétés syntaxiques), à un critère essentiellement sémantique ou à un critère typographique ? Lequel de ces critères faut-il privilégier ? Et surtout lequel est applicable et peut être déterminant dans une perspective de TAL ?

D'après le traitement proposé par les différents outils que nous avons vu, le critère formel semble être celui auquel ces derniers ont recours dans la mesure où il peut être aisément formalisé. Le critère typographique peut se révéler utile dans la mesure où les propositions incises sont généralement placées immédiatement après le verbe ; le critère sémantique joue quant à lui un rôle déterminant lors de la phase de validation.

Les deux outils que nous allons maintenant présenter utilisent une méthodologie toute autre que celles que nous avons définies précédemment. Ces deux outils ont notamment recours à des ressources qui permettent de résoudre les problématiques que nous venons de soulever dans cette section.

4. Une présentation de deux outils disponibles pour le traitement de séquences figées

Intex

Description

Intex est un logiciel créé par le LADL afin de procéder à l'analyse de corpus d'un volume important et langues différentes. Ce logiciel a été développé par Max Silberztein en *1993*. Intex est donc un environnement linguistique permettant d'analyser morphologiquement et syntaxiquement un texte afin de procéder à divers traitements. Ces traitements peuvent consister à rechercher des séquences de diverses natures telles que des lettres, des lexèmes, ou des catégories morphologiques.

Ce logiciel fournit également des outils pour décrire la morphologie flexionnelle et dérivationnelle, la variation orthographique et terminologique. Le vocabulaire est également décrit qu'il s'agisse de mots simples, de mots composés ou d'expressions figées. Des phénomènes dits « semi-figés » sont également recensés dans le logiciel. L'indexation des mots ou d'expressions figées est possible dans le cadre de l'application. Intex permet également l'accès à des concordanciers ou à des outils permettant l'étude statistique des résultats produits.

Les textes ou corpus ouverts en entrée, les dictionnaires électroniques sur lesquels sont basés les analyses ou les grammaires locales utilisées sont représentées à un moment donné du traitement par des transducteurs à états finis. Les transducteurs à états finis sont des graphes qui représentent un ensemble de séquences en entrée et leur associe des séquences produites en sortie. Un transducteur est un automate à état fini. Un automate à état fini, dit aussi automate fini, est un type particulier de transducteur à état fini. La principale différence qui distingue ces deux procédés consiste dans le fait que le transducteur comporte aussi bien une bande de lecture qu'une bande d'écriture tandis que l'automate à état fini comporte uniquement une bande de lecture et ne permet donc pas de production.

Intex a recours à des expressions régulières pour procéder à la recherche de motifs dans les corpus ouverts en entrée. Les graphes qui représentent visuellement les transducteurs à état fini permettent de présenter de manière plus compacte des expressions régulières visant la recherche de motifs complexes.

Le transducteur d'une grammaire représente des séquences de mots du texte et fournit des informations linguistiques d'ordre syntaxique sur ces séquences. Le transducteur d'un dictionnaire représente des séquences de lettres qui correspondent aux entrées des unités lexicales et fournit des informations lexicales sur ces séquences. Le transducteur du texte représente des séquences correspondant aux mots qui constituent une phrase du texte. Ces trois objets, dictionnaires, texte et grammaires, ont donc recours au même mode de représentation, ce qui facilite donc l'implémentation du programme. La notion de transducteur et d'automate est donc essentielle pour comprendre le fonctionnement d'Intex.

Le fonctionnement d'Intex s'appuie justement sur l'exploitation de trois types de ressources linguistiques. Parmi ces ressources nous retrouvons donc les dictionnaires DELA élaborés par le LADL que nous avons décrits dans la section consacrée aux dictionnaires électroniques. Ces dictionnaires, comme nous l'avons donc vu, recensent aussi bien les mots simples que les mots composés.

Les graphes produits par Intex correspondant donc aux dictionnaires, aux grammaires ou aux textes constituent également une de ces ressources linguistiques et présentent l'avantage de présenter de manière compacte des phénomènes linguistiques tant au niveau orthographique, morphologique, que syntagmatique ou syntaxique transformationnel.

Le troisième type de ressource linguistique sur laquelle repose le fonctionnement d'Intex est constitué par les tables du lexique-grammaire qui sont comme nous l'avons vu des bases de données qui fournissent une description détaillée des phénomènes linguistiques qui sont à la frontière des disciplines de la syntaxe et du lexique.

Fonctionnement

Une fois que l'on a démarré Intex, la première étape de traitement d'un corpus consiste à charger un texte en passant par le menu Text > Open … Avant de pouvoir être chargé dans l'application, le texte doit être prétraité et transformé selon des normes propres à Intex et définies par des transducteurs spécifiques. Le chargement du texte s'accompagne par l'apparition d'informations statistiques et formelles concernant ce texte. A partir de ces informations, il est donc possible de voir le nombre de phrases qui composent le texte et également celui des tokens. Il existe quatre sortes de tokens qui sont en fait des objets de base de l'analyse par Intex. Les tokens peuvent représenter les *formes simples* qui apparaissent dans le texte. Les formes simples sont à distinguer des mots simples. Les formes simples sont des séquences de lettres enclavées par deux délimiteurs. Des *tags* qui sont un autre type de tokens représentent des données linguistiques et sont notées entre deux crochets (« { », «} »). Les *digits*, troisième type de tokens, correspondent aux chiffres de 0 à 9. Les délimiteurs, dernier type de tokens, représentent des caractères autres qu'une lettre, un chiffre ou un espace. Ces informations statistiques sont archivées dans un ficher nommé « result.rtf » qui figure dans le répertoire courant.

Outre ces informations statistiques, il est aussi possible d'obtenir des informations d'ordre linguistique sur la nature des formes qui composent le texte. Intex reconnaît en effet quatre types d'unités lexicales : les affixes qui sont des morphèmes dérivationnels ou flexionnels (préfixes ou suffixes pour le français), les mots simples, les mots composés (autrement dit les séquences constituées de plusieurs mots simples) et des expressions figées. Dans le cadre où apparaissent ces informations figure aussi une indication sur la notion d'ambiguïté. Une forme sera effectivement considérée comme étant ambiguë quand deux entrées des dictionnaires DELA correspondent à cette même forme. La notion de token permet justement d'éclaircir ce point. Le token correspond à la forme graphique que prend un mot simple qui figure dans le dictionnaire DELAS. Au mot simple « the » correspondent les trois tokens suivants : « the », « The », et « THE ». Mais seule une entrée de dictionnaire représentent ces trois tokens à savoir l'entrée « the, determiner ».
Cette distinction entre les différentes unités d'analyse d'Intex est importante dans la mesure où les transducteurs ont recours à ces données du texte.

Une fois que le texte a été chargé, il est possible de procéder à une recherche dans ce texte. La fenêtre qui apparaît en cliquant sur la fonctionnalité « Locate Pattern… » (Menu Text) permet de prendre en compte plusieurs paramètres qui permettent d'affiner la recherche.



Figure 4.a) : Intex : Boite de dialogue pour la recherche de motifs

Le motif de recherche apparaît sous la forme de lien hypertexte dans la fenêtre contenant le texte une fois que la recherche a été lancée. Une boîte de dialogue apparaît également à la fin de la recherche : cette boîte de dialogue qui porte l'entête « Display indexed sequences » permet de construire une concordance du motif de recherche. La construction d'une concordance pour une séquence donnée consiste à élaborer une liste des occurrences de cette séquence en affichant également son contexte.

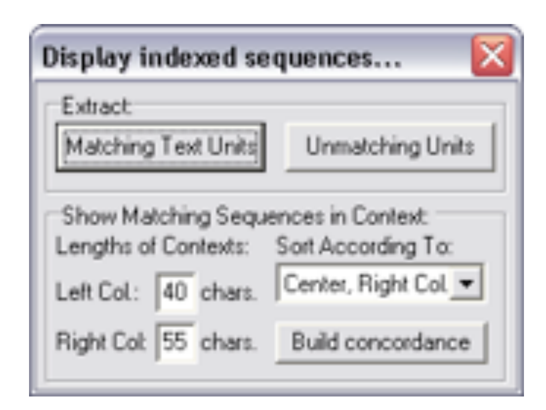


Figure 4.b) : Intex : Boîte de dialogue permettant de construire la concordance du motif recherché

Comme nous pouvons le voir sur cette figure, cette boîte de dialogue offre la possibilité de paramétrer le nombre d'éléments pouvant apparaître dans le contexte gauche ou le contexte droit du motif recherché.

Les diverses options de cette fenêtre représentée par la figure 4.a) permettent notamment de rechercher les formes fléchies d'un mot avec un unique motif de recherche. Il serait donc possible de filtrer toutes les formes fléchies d'un verbe à partir d'un unique motif de recherche ce qui peut s'avérer fort efficace dans le cas des verbes du troisième groupe qui présentent de nombreuses irrégularités :

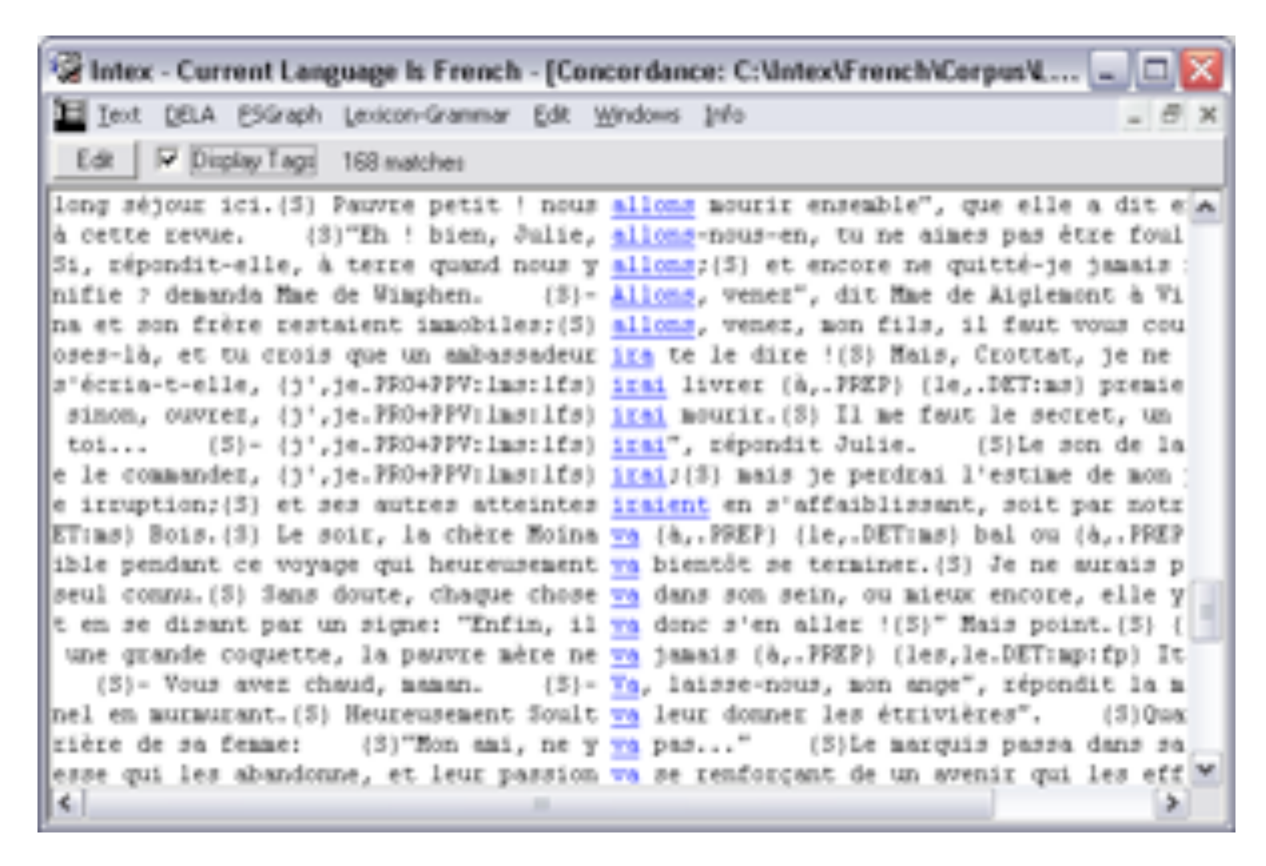


Figure 4.c) : Intex : Filtrage de toutes les formes du verbe aller à partir du motif de recherche <aller>

La figure 4.c) représente le résultat produit par Intex pour la recherche des formes fléchies d'un des verbes les plus irréguliers du français, le verbe « aller ». Le seul motif <aller> permet donc de filtrer toutes les occurrences de ce verbe. Ce type de recherche simplifiée est possible notamment grâce à l'utilisation du dictionnaire électronique DELAF (Dictionnaire Electronique du LADL des Formes fléchies). La possibilité de procéder à ce type de recherche permet de filtrer des formes correspondant à des patrons morphosyntaxiques précis et de construire leur concordance respective à l'issue de la recherche. Cette fonctionnalité permettrait donc de rechercher de manière simplifiée des locutions verbales dans un corpus donné.

Il est également possible de voir certains caractères tels que « {S} » qui résultent de l'opération de prétraitement du texte. Le fait de cocher la case « Display tags » permet d'afficher ces caractères.

Ce type de recherche qui offre pourtant de nombreuses possibilités ne fait pas appel à l'utilisation de transducteurs, du moins cette utilisation se fait en arrière plan. Le recours à des transducteurs s'imposera pour la recherche de motifs complexes. Intex offre donc un outil très

performant et puissant pour ce qui est de la recherche d'informations dans un texte. Une autre fonctionnalité d'Intex tout aussi utile consiste dans le traitement de corpus.

La phase de traitement est précédée par une phase de préparation du texte qui s'opère par l'intermédiaire d'une fenêtre qui porte l'entête « Preprocessing a Text » lors du chargement du texte. Cette fenêtre permet d'appliquer les transducteurs correspondant à la langue du texte et les diverses ressources linguistiques telles que les dictionnaires électroniques DELA.

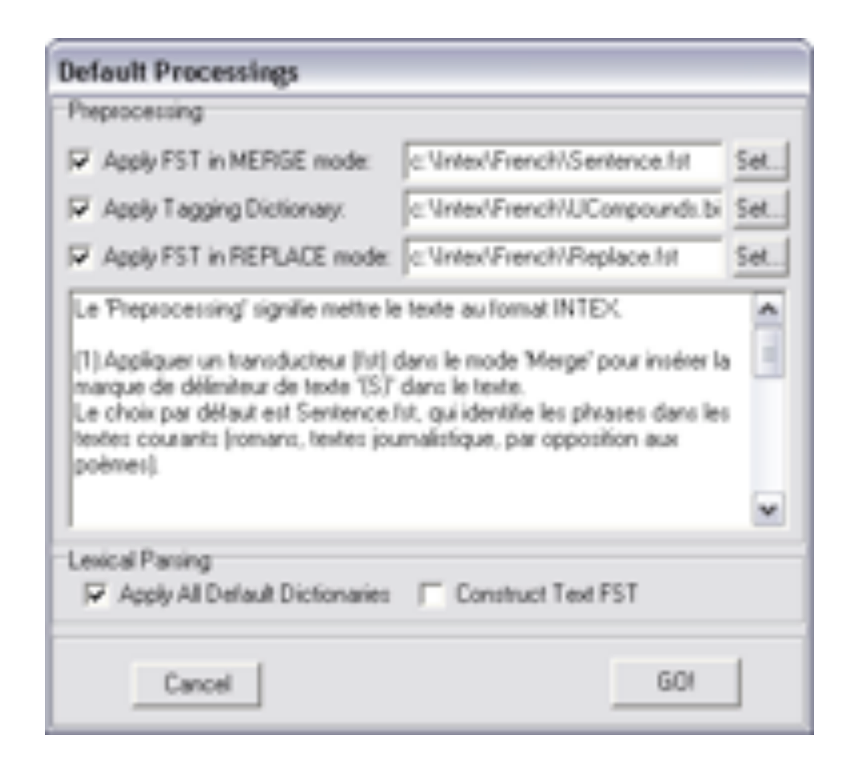


Figure 4.d) : Intex : fenêtre de prétraitement

Un des transducteurs appliqués au texte permet de segmenter le texte. Il s'agit du transducteur « sentence.fst ». Le graphe représenté dans la figure ci-dessous est au format « .grf ». Le transducteur « replace.fst » permet de remplacer les délimiteurs superflus. Cette procédure permet d'insérer des caractères tels que « {S} » qui représentent un retour à la ligne.

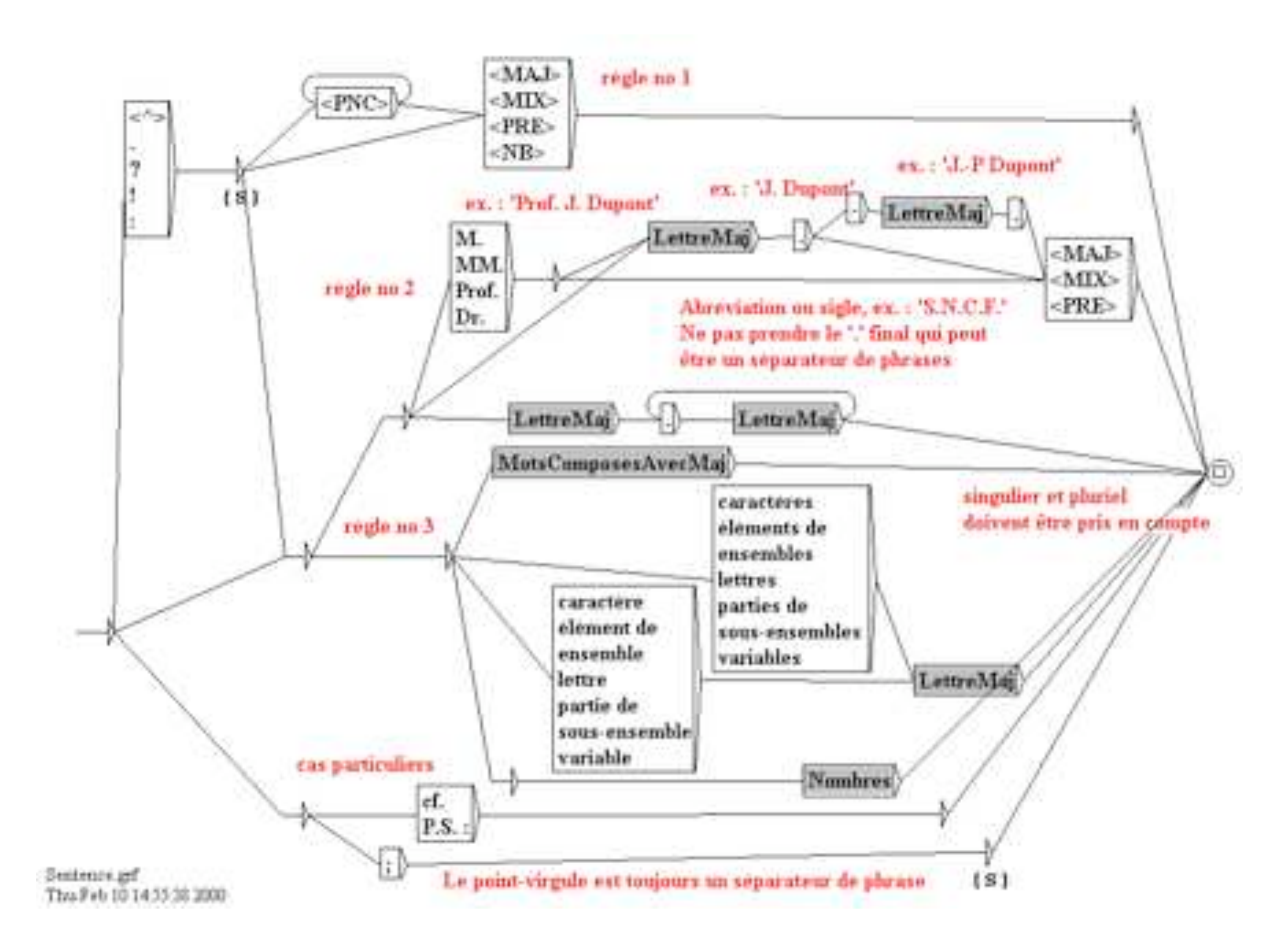


Figure 4.e) : Intex : Graphe « sentence.grf » représentant le transducteur de segmentation du français

Ce graphe décrit donc le mode de segmentation d'une phrase dans un texte donné en langue française.

L'item « Apply lexical Resources » dans le menu « Text » permet de procéder à une analyse lexicale en appliquant donc les ressources lexicales disponibles dans l'application. La boite de dialogue qui apparaît comporte trois zones distinctes dédiées respectivement aux mots simples, aux mots composés, et aux expressions figées. Dans ces zones apparaissent donc les outils, dictionnaires et transducteurs, permettant de procéder à une analyse morphologique. Les dictionnaires électroniques DELAF ou DELACF sont ceux qui sont utilisés pour cette opération ; les dictionnaires qui apparaissent dans les zones indiquées ont une extension « .dic » ou une extension « .bin » : les premiers fichiers peuvent être modifiés, les seconds ne le peuvent pas.

La zone qui nous intéresse le plus est bien entendu celle qui concerne les expressions figées. Les transducteurs lexicaux qui figurent dans cette zone permettent donc de représenter chaque expression figée sous forme de graphe. Ce graphe désigne une expression figée dans sa structure de base ainsi que toutes les variantes qu'elle connaît. Les fichiers qui ont une extension « .fst » correspondent à ces transducteurs. Les tables répertoriées Cxxx correspondent au lexique-grammaire des expressions figées : les noms de fichiers qui ont pour nom Cxxx correspondent donc aux noms des tables du lexique-grammaire. Intex a notamment recours à la table du lexique-grammaire notée C1d.xsl.

Intex permet également de procéder à une analyse syntaxique dont l'objectif premier est la désambiguïsation lexicale. Une boite de dialogue (« Text > Desambiguisation ») permet de procéder à la désambiguïsation au moyen du dictionnaire électronique « disamb.dic ». Intex a recours à des grammaires locales pour éliminer toute ambiguïté lexicale. Les grammaires locales sont des représentations par automate de structures linguistiques complexes qui ne sont pas formalisables par les tables du lexique-grammaire ou les dictionnaires électroniques. Ces grammaires locales sont représentées par des transducteurs et se présentent visuellement sous la forme de graphes.

La construction du transducteur du texte (« Text > ConstructFST-Text ») permet de réaliser l'analyse syntaxique du texte. Cette fonctionnalité génère la construction de transducteur pour chaque phrase du texte. Le transducteur de texte d'une phrase comportant une expression figée telle que « Pierre a perdu la raison » représente donc sous forme de graphe l'analyse syntaxique de cette phrase. Les différents dictionnaires et transducteurs sont appliqués afin de normaliser, désambiguïser et analyser lexicalement le texte.



Figure 4.f) : Intex : Boite de dialogue précédant la construction de l'automate du texte

Cette boîte de dialogue permet donc à l'utilisateur de choisir quelles ressources appliquer pour construire l'automate du texte. Le dictionnaire des expressions figées qui correspond au DELAE est aussi appliqué au texte ainsi que le dictionnaire de désambiguïsation.

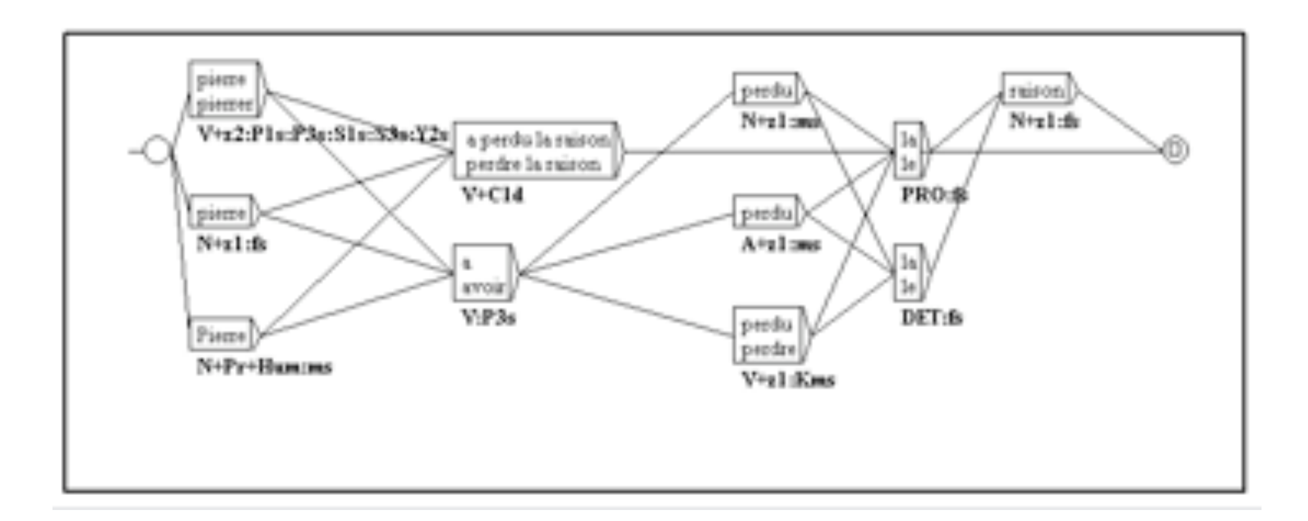


Figure 4.g) : Intex : Graphe représentant le transducteur correspondant à l'analyse de la phrase « Pierre a perdu la raison»

Il est possible de voir sur ce graphe que deux analyses sont possibles pour cette même phrase. La première analyse correspond à une lecture figée dans la mesure où le verbe correspond à « perdre la raison», la seconde analyse représente la lecture compositionnelle. Le fait de représenter cette locution dans une même boite permet de signifier à l'application que celle-ci fonctionne comme un bloc.

Intex dispose par ailleurs d'un éditeur de graphes permettant de formaliser des motifs de recherches ou autres règles. Ces graphes sont générés au format « .grf ». Un module de conversion permet ensuite de transformer le graphe en transducteur à état fini qui porte l'extension « .fst ». L'utilisateur a donc la possibilité de créer ses propres transducteurs afin d'adapter les fonctionnalités proposées par Intex pour le type d'analyse qu'il souhaite faire. L'utilisateur peut donc créer ses propres ressources en fonction de ses besoins.

Intex propose donc de nombreuses fonctionnalités et possibilités en matière de traitement de corpus en procédant aussi bien à une analyse lexicale qu'à une analyse syntaxique. Cet outil est donc d'autant plus puissant et performant qu'il permet de filtrer et d'analyser les expressions figées.

Unitex

Description

Unitex est un logiciel permettant d'analyser et de traiter des corpus d'un volume important. Ce logiciel a été développé par Sébastien Paumier à l'Institut d'Electronique et d'Informatique Gaspard Monge de l'Université de Marne-la-Vallée. Cette application est un ensemble de logiciels qui s'appuie sur des ressources linguistiques pour traiter des textes dans des langues naturelles différentes. Ces ressources se présentent sous la forme de dictionnaires électroniques comme nous l'avons vu dans la section qui leur était consacrée. Ces dictionnaires électroniques recensent de manière exhaustive les différentes formes

linguistiques disponibles dans une langue. Les Laboratoires RELEX qui constituent un réseau informel de laboratoires ont travaillé dans la mise à disposition de cet outil pour des langues d'usage « courant » telles que le français, l'anglais ou l'espagnol mais également pour des langues aussi « exotiques » ou peu courantes que le thaï ou le norvégien. Ce logiciel fonde également son fonctionnement sur l'exploitation de ressources linguistiques. Unitex a donc également recours à l'utilisation de transducteurs à état fini pour traiter les corpus ouverts en entrée. Les dictionnaires électroniques DELA et les tables du lexique-grammaire élaborées par le LADL sont également utilisés par ce logiciel.

Tout comme le logiciel Intex, Unitex utilise des tokens comme unités d'analyse : nous retrouvons donc les formes simples, les digits et les délimiteurs auxquels Intex a recours pour procéder à son analyse.

Fonctionnement

Au lancement du logiciel, une boite de dialogue propose de choisir le répertoire de travail dans lequel l'utilisateur souhaite travailler. Toutes les ressources résultant du traitement de corpus sont enregistrées dans ce répertoire. L'utilisateur est ensuite amené à choisir la langue sur laquelle il souhaite travailler : l'anglais, le français, le grec, le norvégien, l'italien, le portugais, le russe et le thaï.

Le chargement d'un texte s'accompagne dans ce logiciel d'une boite de dialogue permettant de procéder à un prétraitement du texte. La fenêtre de prétraitement portant l'entête « Preprocessing And Lexical Parsing » se présente sous la forme suivante :

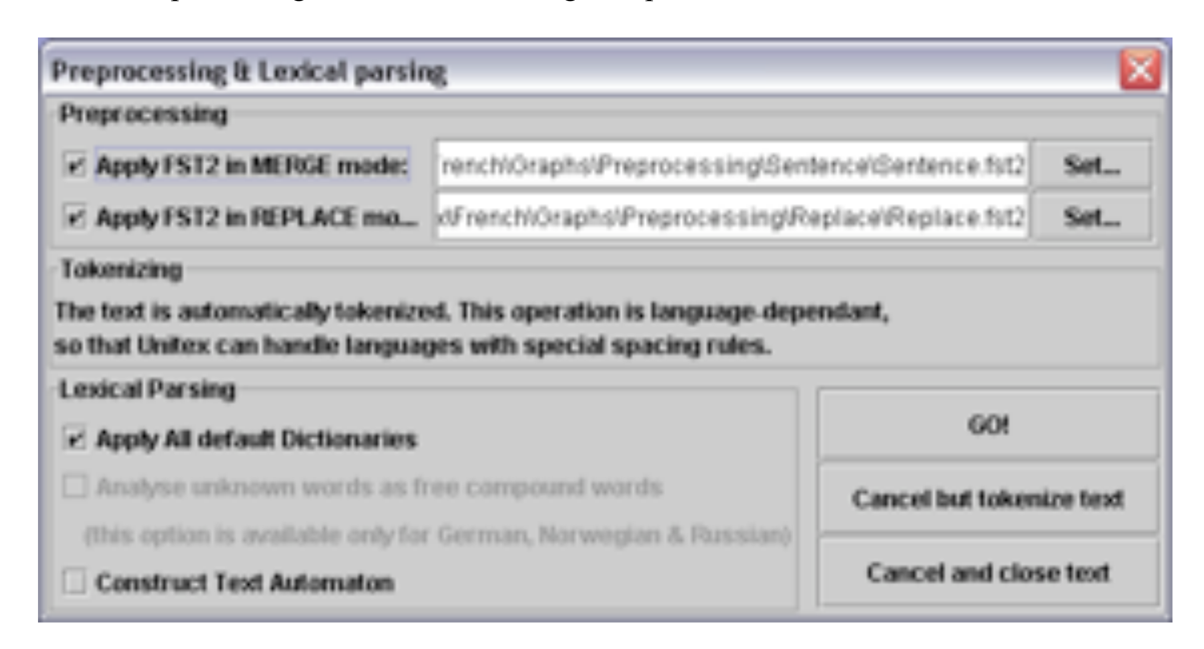


Figure 4.h) : Unitex : Fenêtre permettant de procéder au prétraitement du texte et à l'application des ressources linguistiques

Comme nous pouvons le voir, cette fenêtre est proche de celle proposée par Intex. La première option permet de segmenter le texte en phrases, la seconde permet d'effectuer des

remplacements dans le texte. Ces remplacements consistent essentiellement à procéder à une normalisation du texte. Cette normalisation vise à remplacer les séparateurs selon deux principes. Ce premier principe implique que les suites de séparateurs comportant au moins un retour à la ligne soient remplacées par un unique retour à la ligne. Le second principe suggère que toute autre suite de séparateurs soit remplacée par un espace. D'après ces principes, seuls les espaces et les retours à la ligne constituent des séparateurs pertinents dans une analyse linguistique. Cette gestion des séparateurs s'explique par le fait que l'importance occupée par ces derniers diffère selon les langues. C'est notamment le cas pour les langues asiatiques qui interdisent, obligent ou rendent facultative la présence d'espaces. Il est à noter que le fichier prétraité comporte une extension « .snt » de sorte que le fichier original n'est pas altéré. Ce même type d fichier est généré par Intex lors du traitement de corpus.

Le fichier « Sentence.fst2 » comporte la grammaire qui permet de segmenter le texte en français.

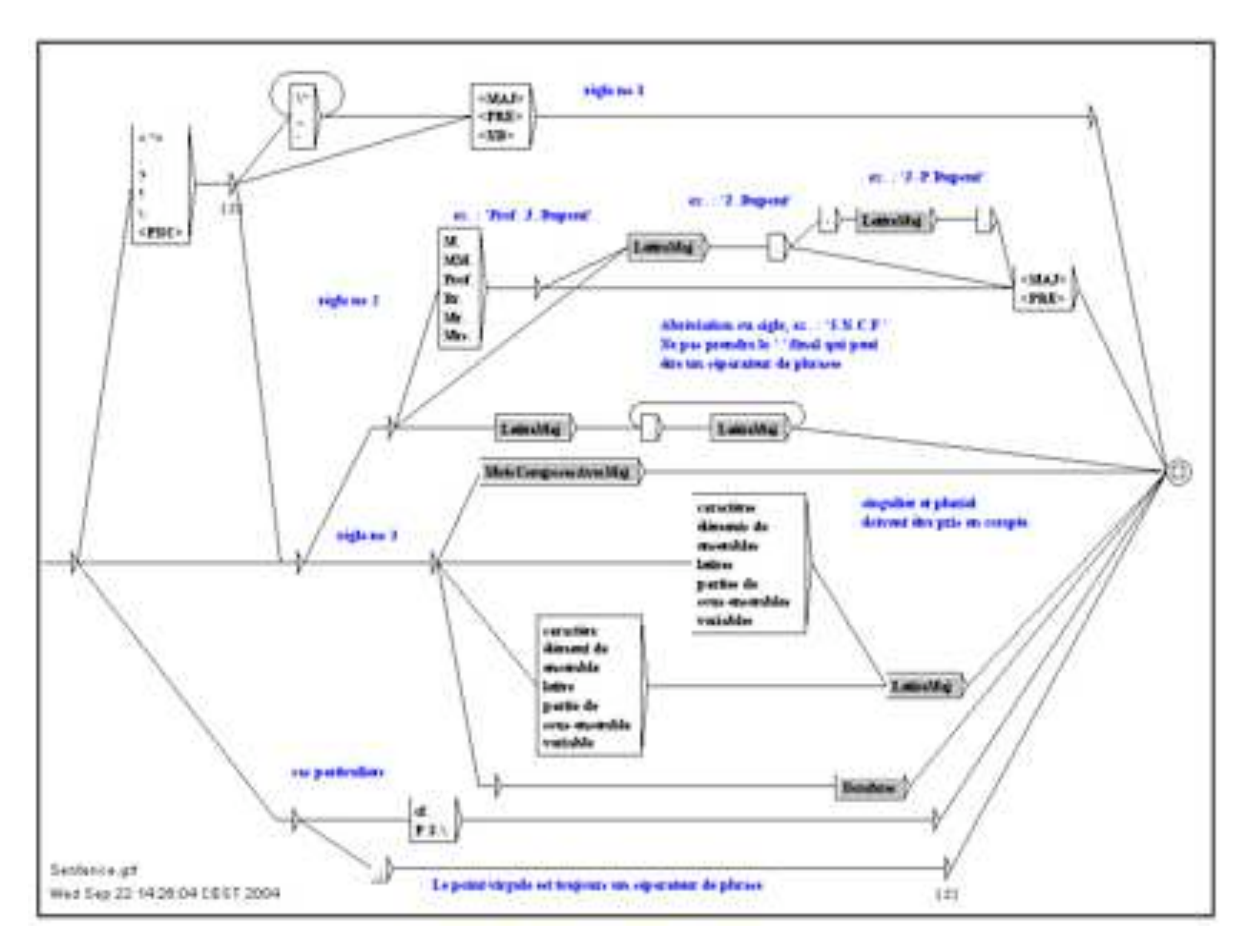


Figure 4.i) : Unitex : Graphe représentant la grammaire de segmentation du français

Ce graphe permet donc de transcrire visuellement que la présence de signes de ponctuation suggère que les séquences qui suivent constituent des phrases. Ce graphe permet également d'indiquer que la présence d'un « point » n'est pas nécessairement le signe d'une fin de phrase dans la mesure où le point sert aussi marquer des initiales et des abréviations comme dans l'exemple suivant « `M. Dupont` ».

Unitex distingue pour le français trois types d'unités lexicales qui sont donc le séparateur de phrase noté « {S} », une suite de lettres ou tout caractère qui n'est ni une lettre, ni un séparateur de phrase. Le dernier type d'unité lexicale peut donc désigner un espace.

Une fois que le texte est chargé et prétraité, il est possible d'obtenir des informations statistiques et linguistiques dans la même fenêtre que le texte. Deux autres fenêtres dans lesquelles les différentes formes qui apparaissent dans le texte sont triées et répertoriées. Une de ces fenêtres où sont visibles trois zones distinctes permet notamment de voir dans chacune de ces zones une liste de mots simples, une liste de mots composés ainsi qu'une liste de mots simples inconnus qui figurent dans le texte. La présence de ces deux fenêtres résulte de l'application des dictionnaires et des ressources linguistiques. Parmi la liste des mots inconnus figureront donc les noms propres car ils ne figurent ni dans le dictionnaire électronique de mots simples ni dans celui des mots composés.

Unitex permet également de procéder à des recherches poussées sur le texte. Une boite de dialogue (qui apparaît en cliquant sur l'item « Text > Locate Pattern ») permet de paramétrer cette recherche détaillée selon un certain nombre d'options comme nous pouvons le voir dans la figure ci-dessous :



Figure 4.j) : Unitex : Boite de dialogue permettant de procéder à une recherche

Les résultats produits sont présentés sous la forme de concordance dans un fichier nommé « concord.html » qui est ouvert dans une autre fenêtre que celle où figure le texte. La figure ci-dessous représente la boite de dialogue qui permet d'afficher les occurrences recherchées et de construire la concordance du motif recherché :



Figure 4.k) : Unitex : Boite de dialogue permettant de construire la concordance du motif recherché

Nous pouvons voir sur cette figure que cette boite de dialogue est très ressemblante avec celle proposée par Intex. Cette fenêtre propose cependant quelques options supplémentaires comme une optimisation de l'affichage des concordances en utilisant un navigateur web. Il est également possible de sauvegarder le résultat de la recherche dans un fichier de sortie.

Comme nous l'avions fait en utilisant Intex, nous avons saisi le motif de recherche <aller> pour sélectionner toutes les formes fléchies de ce verbe.



Figure 4.l) : Unitex : Concordance du motif de recherche <aller>

Nous pouvons donc voir que les résultats produits sont similaires à ceux produits par Intex.

Nous pouvons également voir sur cette figure qu'Unitex n'a pas recours aux mêmes tags lors de l'application du transducteur de remplacement au moment du prétraitement du texte.

Un automate du texte peut être généré en cliquant sur « Construct FST-Text » dans le menu « Text » ce qui permet de procéder à l'analyse syntaxique du texte. Comme nous avions procédé dans la section consacrée à l'étude d'Intex, nous avons construit l'automate du texte pour la phrase « Pierre a perdu la raison » qui est une locution verbale.

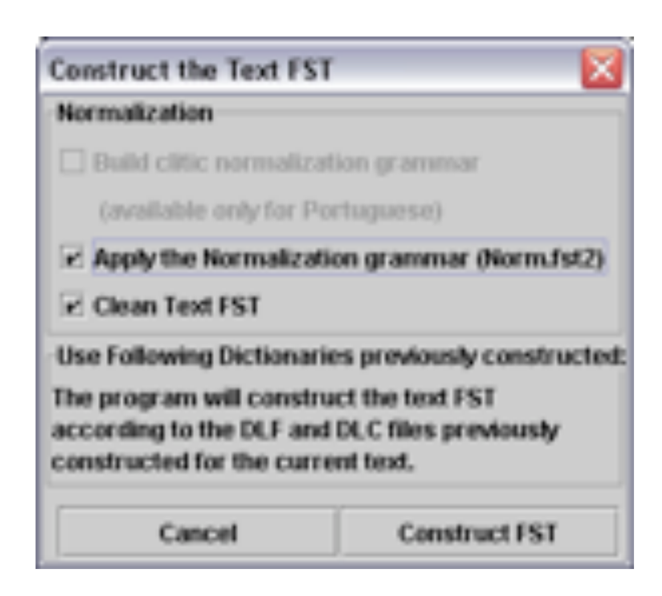


Figure 4.m) : Unitex : Boite de dialogue permettant de construire l'automate du texte

Un transducteur de normalisation est appliqué afin de construire le transducteur du texte.

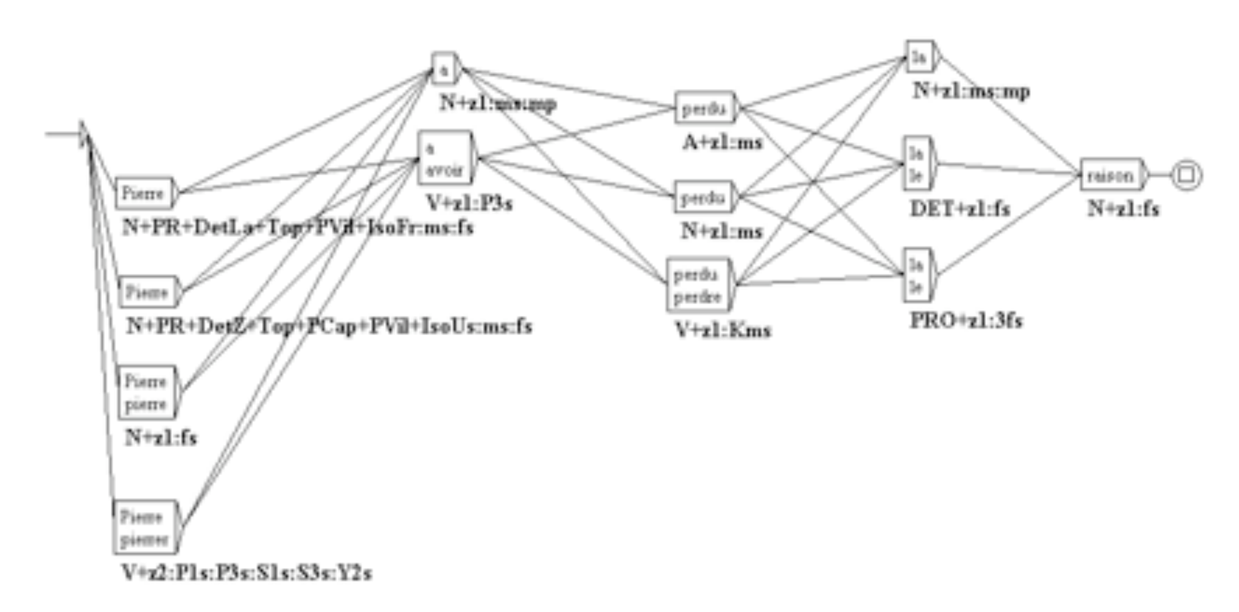


Figure 4.n) : Unitex : Automate du texte de la phrase « Pierre a perdu la raison »

Nous pouvons donc voir sur ce graphe que la séquence « perdre la raison » n'est pas considérée comme une séquence composée ou comme une expression figée malgré

l'application des tables du lexique-grammaire et des dictionnaires DELA. La boite de dialogue indique en effet que seuls les dictionnaires DELAC et DELAF sont appliqués. La séquence « perdre la raison » n'apparaît pas dans une même boite ce qui signifie que seule la lecture compositionnelle est permise.

Unitex dispose par ailleurs d'un éditeur de graphes qui permet à l'utilisateur de créer ses propres ressources. Il serait donc possible d'envisager la création de grammaires locales visant l'extraction de certains types de séquences figées. Unitex offre de plus la possibilité d'importer des graphes générés par Intex.

Cet outil, malgré sa forte ressemblance avec les fonctionnalités et l'interface d'Intex, est tout de même différent d'Intex. En effet, seuls les termes composés nominaux ou adverbiaux sont pris en compte. Les locutions verbales ne peuvent donc pas être analysées avec Unitex.

Nous pouvons voir au travers de l'étude de ces deux outils qu'Intex et Unitex proposent des fonctionnalités similaires et efficaces pour le traitement de corpus. La méthodologie utilisée par ces deux logiciels est fondamentalement la même ainsi que les ressources utilisées. Cette méthodologie diffère totalement de celles que nous avons définies dans la section précédente mais ne s'avère pas moins efficace.

5. Application sur un corpus

5.1 Constitution d'un corpus

Le corpus qui a été utilisé dans le cadre ce travail a été constitué à partir de la version électronique du quotidien Le Monde. Ce fichier qui correspondait à la version du 13 avril 2003 était au format XML. Ce fichier a ensuite été modifié afin de constituer un fichier texte au format brut. Le choix d'un corpus journalistique peut se révéler « périlleux ». En effet, ce choix implique un autre type de contrainte et de difficulté : les phénomènes de défigement. Les défigements peuvent être sémantiques et / ou syntaxiques. Ce phénomène est assez courant et concerne autant la langue écrite que la langue parlée. La presse française a très souvent pour ne pas dire systématiquement recours au défigement pour produire les unes et les articles les plus accrocheurs possibles. La superposition d'une lecture figée à une lecture littérale provoque généralement une ambiguïté souvent recherchée par les médias pour créer un effet humoristique.

Ex 5.1.a) : Tyler Hamilton, trahi par son sang.

L'exemple 5.1.a) est bien un exemple de défigement. Tyler Hamilton, cycliste, a été contrôlé positif à un test de dopage. L'expression « trahi par son sang » implique généralement un emploi au « sens figuré » (lecture figée) : le mot « sang » fait dans cet emploi là référence aux liens familiaux. Cette expression est ici employée au « sens propre » (lecture compositionnelle).

Selon Alain Rey (*1997*), le défigement suppose « des modifications dans un arrangement stable supposé connu, mais non pas dans l'effet global – sémantique et

pragmatique – de l'unité phraséologique considérée ». Cette définition du défigement sous-entend que les modifications que connaît une séquence figée pour subir un défigement sont minimes : il est donc toujours possible de reconnaître la structure de base de la locution.

François Rastier (*1997*) précise quant à lui que « les défigements témoignent de l'incidence du contexte sur la lexie, et plus généralement du global sur le local. » les défigements dont parle Rastier sont des défigements par contexte.

Il existerait donc deux types de défigements : dans le premier cas, la modification de la structure de la locution provoque le défigement ; dans le second, c'est l'emploi d'une locution dans un contexte qui ne s'y prête pas qui crée le défigement.

Ce corpus est donc constitué de deux articles extraits de cette édition électronique du 13 avril 2003. Le premier article intitulé « Après la guerre, la récession mondiale n'aura pas lieu » qui traite des conséquences politiques et économiques de la guerre en Irak. Le second article intitulé « Astro Boy est la passion des Nippons pour les humanoïdes » aborde un sujet plus léger qui est celui d'un dessin animé japonais Astro Boy qui connut un grand succès dans les années 1970-80. Ces deux articles ne figuraient pas en une et ont été choisis en fonction de leur longueur relative et des thèmes abordés.

5.2 Analyse du corpus

Avant de procéder à une analyse du corpus par des outils régulièrement utilisés par le TAL tels que Intex ou Unitex, nous avons procédé à une analyse manuelle. Nous avons donc analysé les formes verbales une à une. Les critères que nous avions décrits et illustrés en 2.3 (Traits caractéristiques des locutions verbales) ont été repris pour analyser ce corpus.

Les tableaux ci-dessous récapitulent et illustrent ces critères.

Ex : Prendre la tangente ⇨ Pierre a pris la tangente		
	Transformation	***Exemples***
1	**Passif**	** La tangente a été prise par Pierre.*
2	**Extraction**	*? C'est la tangente que Pierre a prise.*
3	**Détachement**	*? La tangente, Pierre l'a prise.*
4	**Pronominalisation**	** Pierre l'a prise (la tangente).*
5	**Relativation**	** La tangente que Pierre a prise.*
6	**Interrogation**	** Qu'est-ce que Pierre a pris? La tangente.*

Figure 5.2.a) : Tableau récapitulatif des Critères transformationnels appliqués à la phrase comportant la locution verbale « Pierre a pris la tangente »

N0	Groupe Nominal Sujet
N1	Groupe Nominal Objet
N2	Groupe Nominal Objet Second
V	Verbe ou participe passé
Aux	Auxiliaire

Figure 5.2.b) : Abréviation employée pour formaliser les structures syntaxiques (d'après Maurice Gross)

Le tableau ci-dessus définit quelles sont les abréviations que nous allons utiliser pour formaliser ces structures syntaxiques figées.

Nous avons tenté dans la figure représentée ci-dessous une formalisation de ces critères pour appliquer ces transformations à diverses phrases.

	Transformation	***Exemples***
1	**Passif**	*(*) N1 Aux V par N0.*
2	**Extraction**	*(*) C'est N1 que N0.*
3	**Détachement**	*(*) N1, N0 PRON.(N1) V.*
4	**Pronominalisation**	*(*) N0 PRON.(N1) V.*
5	**Relativation**	*(*) N1 QUE N0 V.*
6	**Interrogation**	*(*) Qu'est-ce que N0 V? N1.*

Figure 5.2.c) : Formalisation des structures syntaxiques produites par l'application des critères transformationnels

Nous avons décidé de présenter les résultats de cette analyse sous la forme d'un tableau fortement inspiré par les tables du lexique-grammaire.

Les séquences du corpus apparaissent dans les lignes, le lemme qui pourrait correspondre à la structure de base de cette séquence apparaît dans la colonne suivante. Quand une séquence est sémantiquement opaque, le signe « + » figure dans la colonne correspondante. Le signe « - » figurant dans les colonnes correspondant au critère formel indique que la transformation correspondante n'est pas possible. Le point d'interrogation « ? » indique que le résultat produit par la transformation ne parait pas naturel mais n'est pas agrammatical pour autant. Les « X » indiquent que les séquences décrites n'ont pas été analysées en détail étant donné qu'il s'agit de verbes libres.

Ce tableau d'analyse figurant en annexes indique si les séquences verbales figurant dans le texte sont des verbes libres, des verbes supports ou des verbes entrant dans des séquences figées.

Nous pouvons voir dans ce tableau que ce corpus compte près 13 de locutions verbales. Il est aussi possible de voir que 5 verbes support figurent dans ce texte. Cette classification peut être contestée dans la mesure où d'autres critères sont applicables pour distinguer ces séquences. Cette analyse a été particulièrement difficile en raison des diverses contraintes que nous avons vues dans la section 3 (Contraintes liées aux locutions verbales). En effet, certaines constructions figées emploient des verbes qui ne le sont pas : la question s'est alors posée pour ces cas de déterminer si le verbe joue juste un rôle de support ou s'il peut être considéré comme étant figé. De même que dire des verbes employés avec des locutions adjectivales ou adverbiales ?

Pour la construction du programme Verbalex nous avons adopté la démarche suivante : toute suite verbe - complément(s) est considérée comme étant une locution verbale dés lors que le verbe ou l'un des compléments présentent un degré quelconque de figement.

5.3 Traitement et résultats produits par les logiciels Intex et Unitex

5.3.1 Intex

Le chargement du corpus sous Intex ne nécessite pas un format d'encodage précis pour le fichier ouvert en entrée.

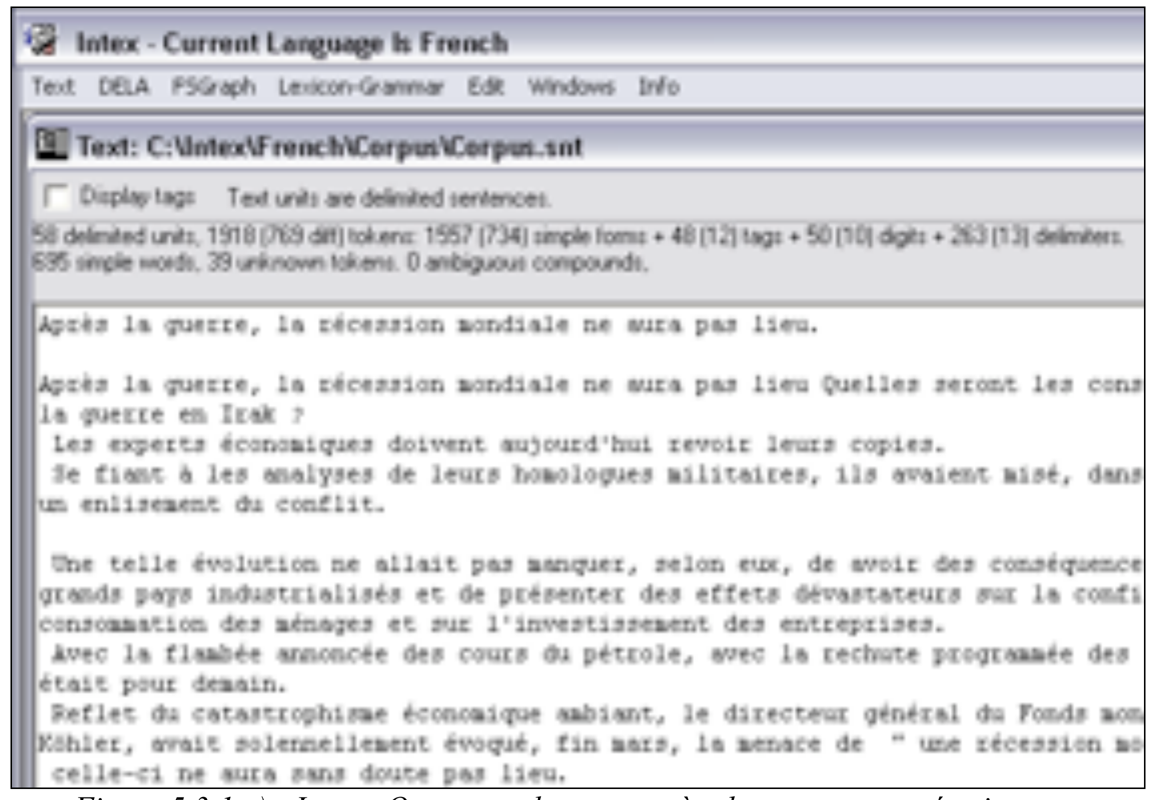


Figure 5.3.1.a) : Intex : Ouverture du texte après chargement et prétraitement.

Il est donc possible de voir apparaître les informations d'ordre statistiques et linguistiques. Il est à noter que le nombre de mots composés n'apparaît pas dans le cadre.

Nous avons après avoir chargé le texte appliqué les ressources lexicales. Certains dictionnaires étaient sélectionnés par défaut : nous avons décidé cependant d'appliquer toutes les ressources disponibles comme nous pouvons le voir dans la figure 5.3.1.b).

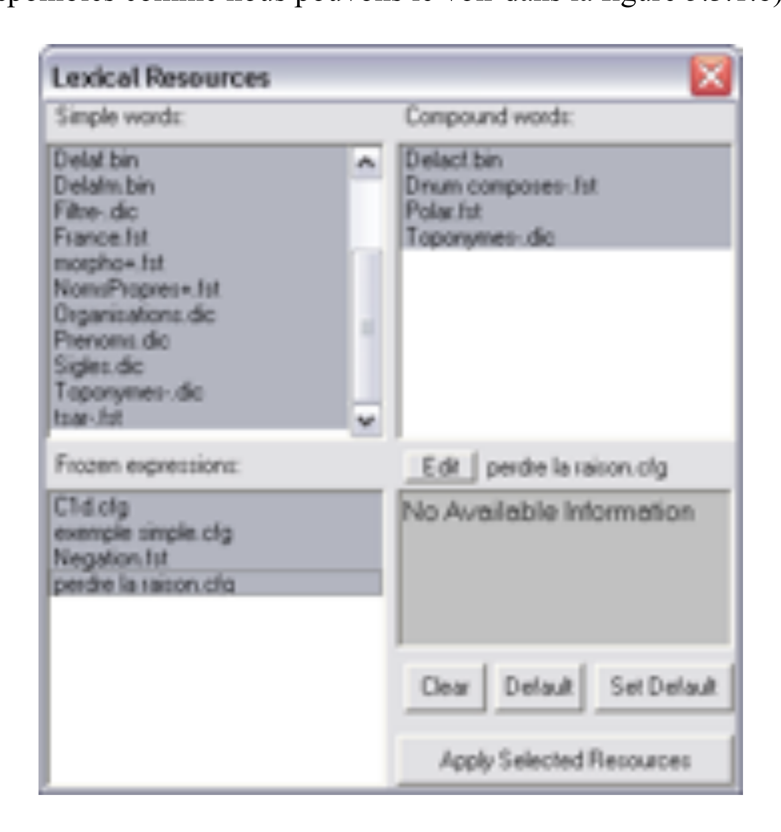


Figure 5.3.1.b) : Intex : Application des ressources lexicales.

L'application des ressources lexicales permet de procéder à l'analyse lexicale du texte. Dans le cadre de ce travail qui porte sur les séquences figées, il est surtout important d'appliquer toutes les ressources disponibles dans la zone correspondant aux expressions figées. La table du lexique-grammaire qui décrit les expressions figées du français correspond au fichier « C1d.cfg ». L'application de cette ressource en particulier devrait donc produire des résultats satisfaisants.

L'application des ressources lexicales (Text > Apply Lexical Resources) permet d'actualiser le dictionnaire des formes apparaissant dans le texte après le prétraitement. L'application de toutes les ressources disponibles dans Intex produit donc le dictionnaire de formes suivant :

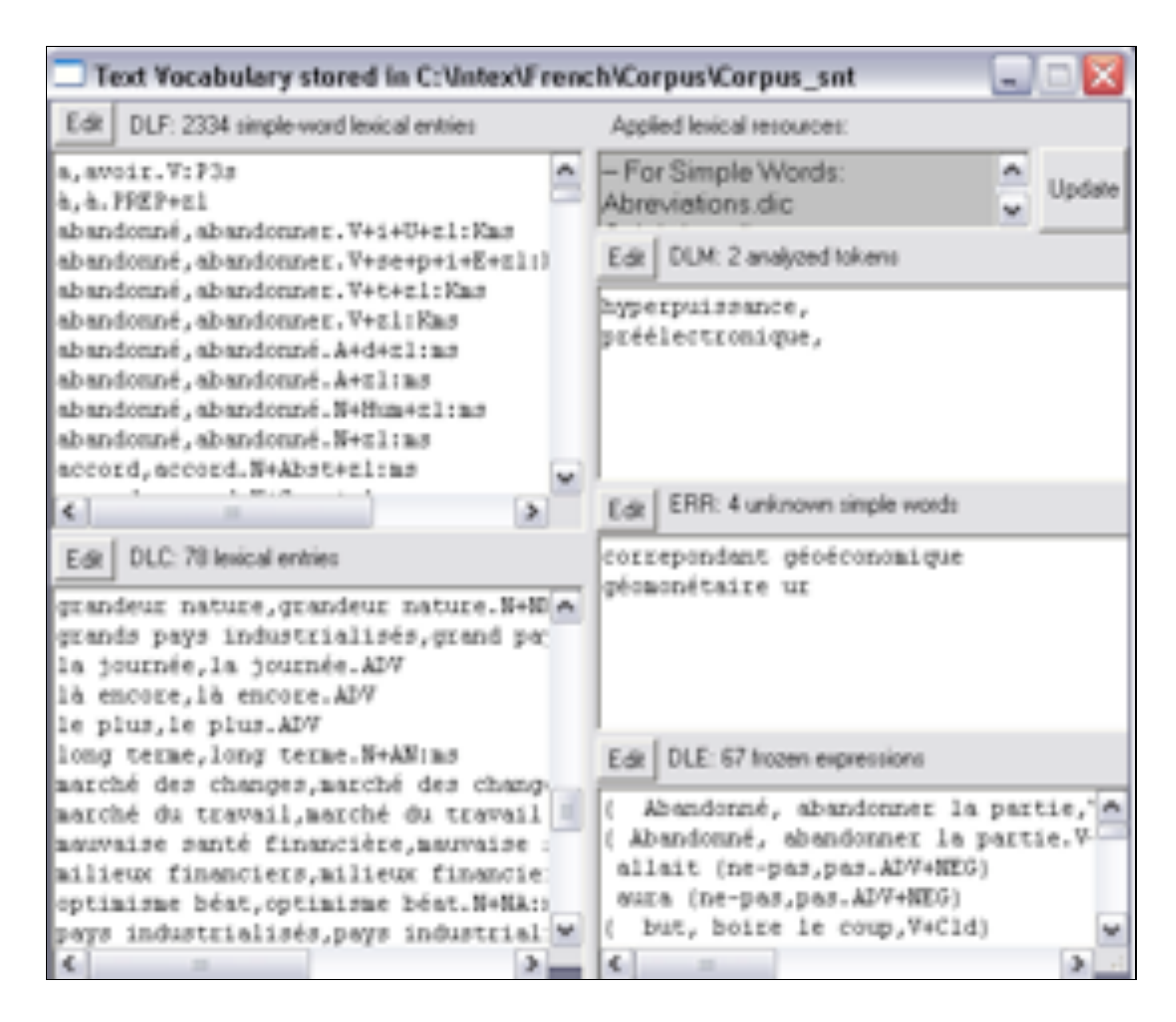


Figure 5.3.1.c) : Intex : Analyse lexicale : dictionnaire des formes du texte

Nous pouvons voir sur cette dernière figure que le logiciel Intex considère qu'il y a 78 mots composés, il y aurait également 67 expressions figées. La liste correspondant à ces dernières séquences figure en annexes. Ce nombre paraît important après l'analyse manuelle à laquelle nous avons procédé. Il est possible de remarquer que figurent sur cette liste les mots qui apparaissent dans la tables du lexique-grammaire C1d : si un mot qui figure dans une expression figée qui est décrite dans la table apparaît dans le texte alors ce mot et ses cooccurrences sont comptabilisés en tant qu'expressions figées.

Le mot « abandonné » n'apparaît qu'une fois dans ce corpus, en l'occurrence au début du second article qui constitue le second paragraphe de ce texte. La construction de l'automate du texte de la phrase « Abandonné par son créateur dans un cirque pour robots, Astro Boy sera accueilli par un autre scientifique avant de embrasser une carrière de super-héros à le service de la paix et de l'harmonie entre les humains et les machines » produit le résultat suivant :

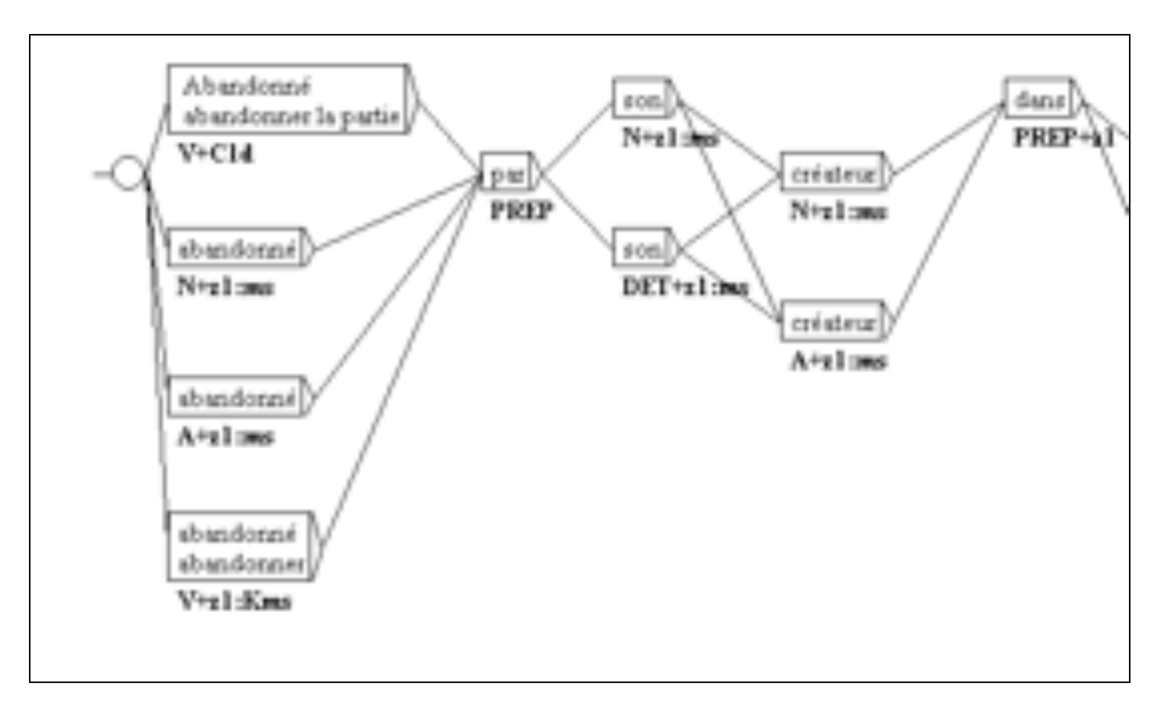


Figure 5.3.1.d) : Intex : Automate du texte de la phrase contenant la séquence « Abandonné »

Cette séquence apparaît donc dans la zone consacrée aux expressions figées car le mot « abandonner » ou une forme fléchie de ce verbe figure dans l'expression figée « abandonner la partie ». Le graphe produit par l'application du dictionnaire de désambiguïsation et par la construction de l'automate du texte indique qu'une des analyses possibles de cette phrase inclurait l'expression figée « abandonner la partie ».

5.3.2 Unitex

Les corpus pris en entrée par Unitex doivent être enregistrés au format Unicode Little-Endian Text. Le corpus nommé « CorpusUnitex.txt » a donc été enregistré au format unicode. L'étape de prétraitement qui suit le chargement du texte permet d'indiquer les informations d'ordre statistiques et linguistiques.

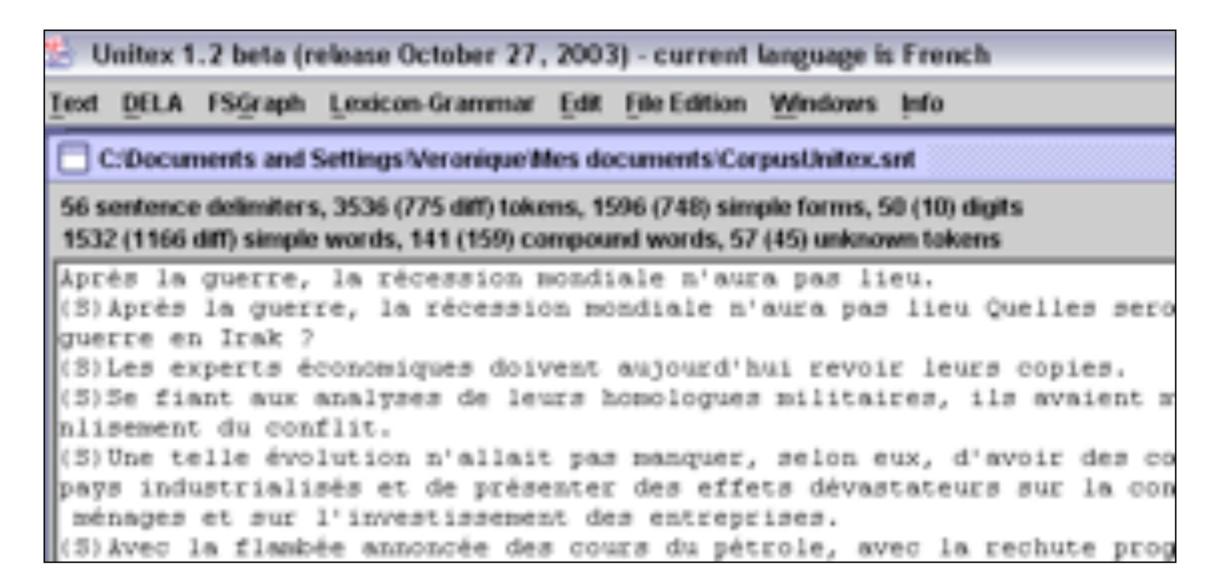


Figure 5.3.2.a) : Ouverture du texte après chargement et prétraitement.

On peut donc voir sur cette figure le nombre de tokens présents dans le texte : il y a donc 3536 tokens, 1532 mots simples, 141 mots composés. Les sauts de ligne ont été remplacés par {S}. Les remplacements ont été effectués après la création d'un fichier nommé « CorpusUnitex.snt » : le fichier original n'a donc pas été altéré.

L'application des ressources lexicales permet d'afficher et de lister les unités lexicales selon qu'il s'agit de mots simples, des mots composés et des mots inconnus, ainsi que les tokens du corpus. La boite de dialogue qui permet d'appliquer les ressources en dehors de l'étape de prétraitement se présente ainsi :

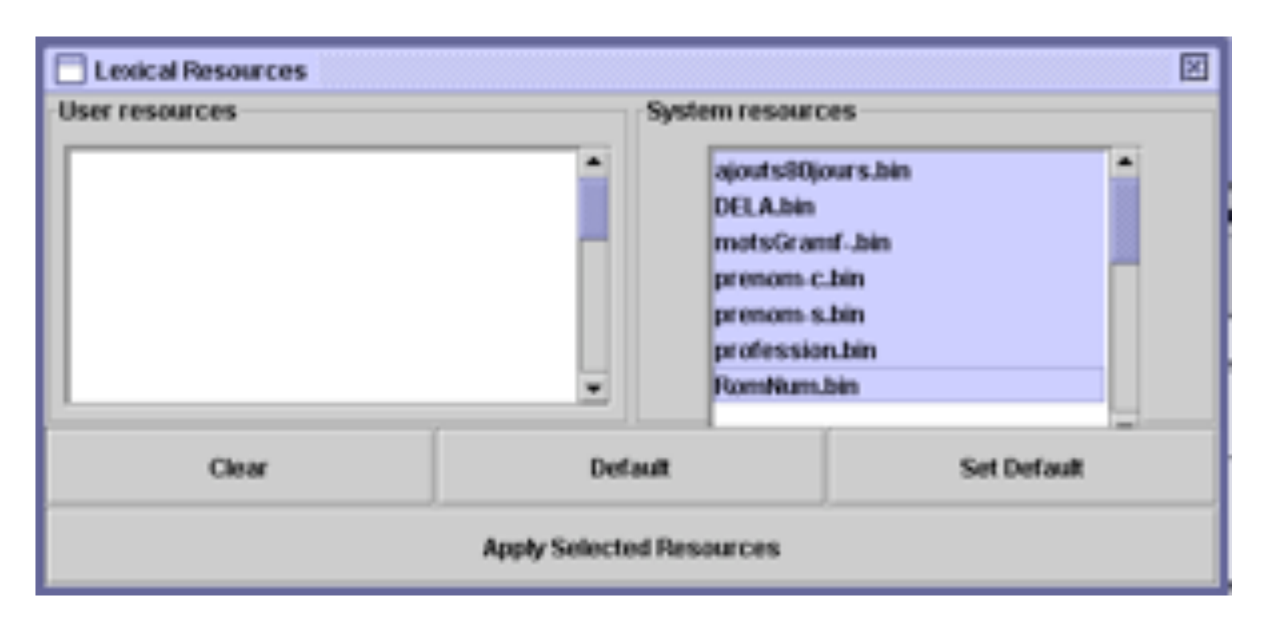


Figure5.3.2.b) : Unitex : Application des ressources lexicales

Cette figure nous montre quelles sont les ressources lexicales disponibles dans Unitex.

L'utilisateur peut en plus des ressources déjà disponibles appliquer des ressources externes au programme ou qu'il a lui-même créées. Nous pouvons d'ores et déjà constater que le nombre de ressources disponibles est moins important que dan Intex.

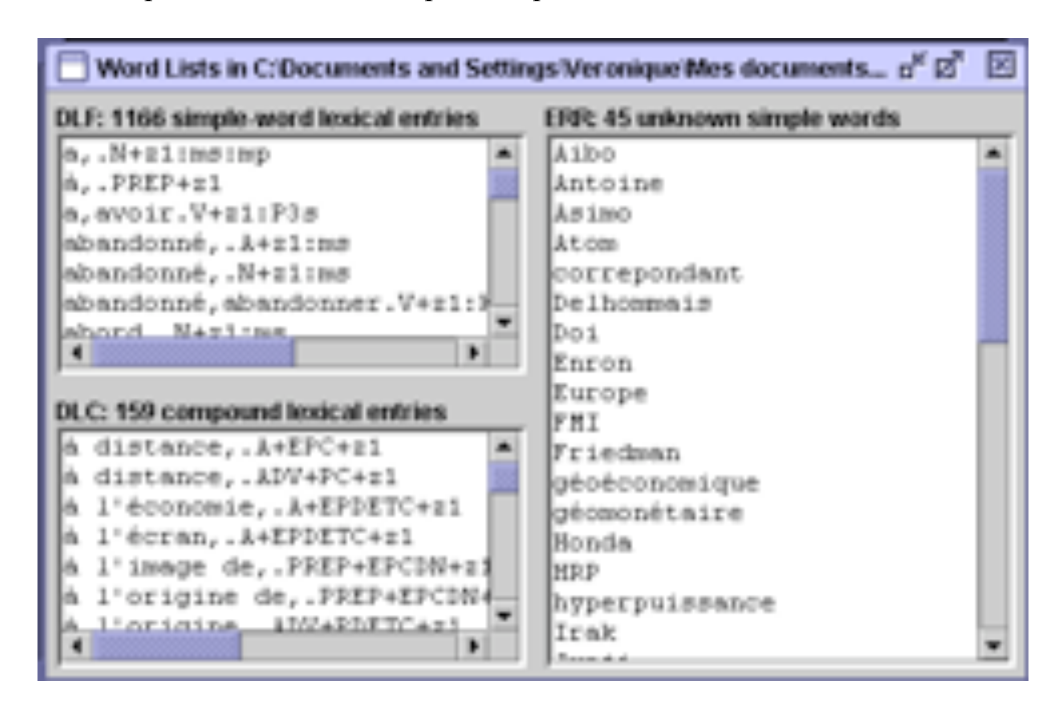


Figure5.3.2.c) : Unitex : Dictionnaire des formes du texte

La liste des mots composés qui apparaissent dans ce corpus, catégorie qui nous intéresse dans le cadre de ce travail, produite par Unitex figure dans les Annexes. Il est possible de voir dans cette liste qu'aucune locution verbale n'apparaît : seules les suites figées de nature nominale, adjectivale ou adverbiale.

La construction de l'automate du texte pour ce corpus permet de voir que les formes de ce texte ne comporte pas de locutions verbales.

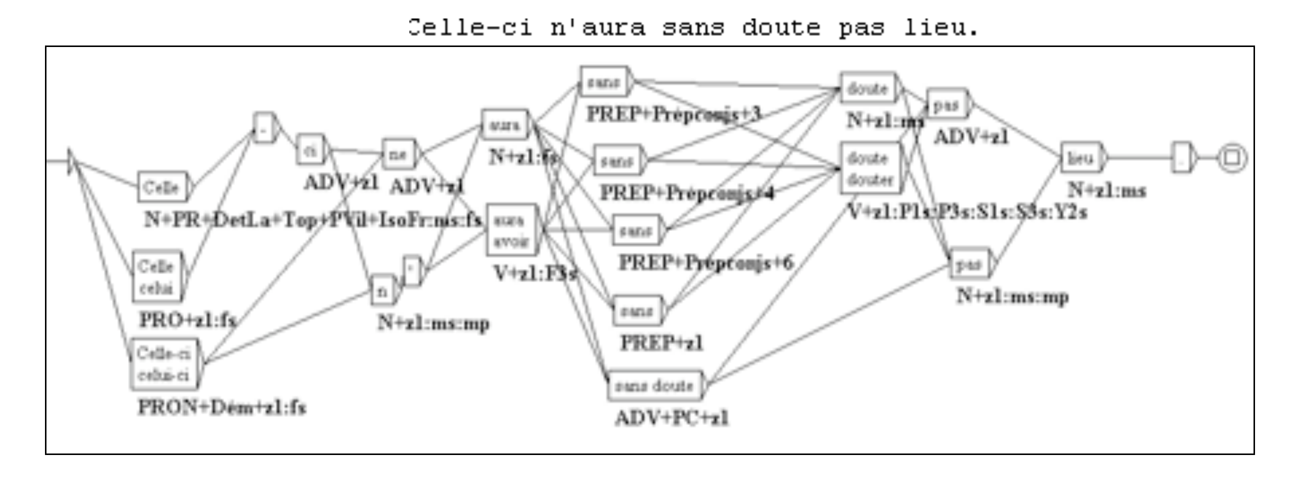


Figure 5.3.2.d) : Unitex : Automate du texte de la phrase « Celle-ci n'aura sans doute pas lieu ».

Il est possible de voir sur ce graphe que la suite adverbiale « sans doute » est bien considérée comme une locution adverbiale par Unitex. Cette locution apparaît également dans la liste des mots composés en annexes.

Au travers de l'application de ce corpus à ces deux outils, il apparaît que l'interface de ces outils intègre la notion d'expressions figées notamment Intex. Le traitement proposé par ces deux logiciels n'est cependant pas assez efficace. La construction de l'automate du texte pour une phrase telle que « Pierre a cassé sa pipe » comportant la locution verbale « casser sa pipe » aussi bien dans Intex que dans Unitex n'est pas considérée comme une locution verbale, autrement dit comme une expression figée.

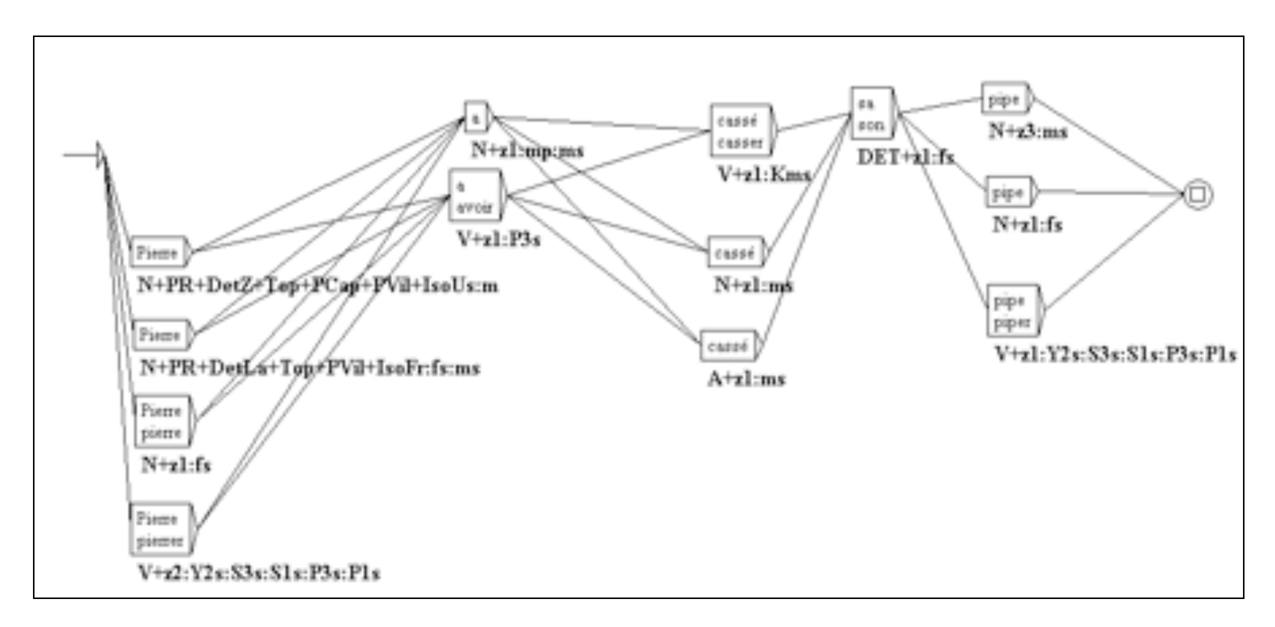


Figure 5.3.2.e) : Unitex : Automate du texte de la phrase « Pierre a cassé sa pipe »

Intex produit le même type d'analyse pour cette phrase qui comporte pourtant une locution verbale. La raison pour laquelle ces deux logiciels ne parviennent pas à reconnaître cette suite comme étant figée réside dans le fait qu'Unitex n'a pas recours à la table du lexique grammaire des expressions figées. En effet, cette table n'apparaît pas dans une forme quelconque dans les ressources disponibles dans l'application. De plus, la zone consacrée aux mots composés ne fait pas apparaître les séquences verbales. En ce qui concerne Intex, ce problème pourrait s'expliquer par le fait que la principale ressource lexicale en matière d'expressions figées est la table C1d mais comme on pourra le voir en annexes, cette table ne décrit que les expressions correspondant au patron syntaxique précis (N0 V (DET) N1) et l'expression « casser sa pipe » ne figure pas dans cette table. En effet, s'il ne fait aucun doute que cette expression est figée, celle-ci est plus difficile à formaliser dans la mesure où le possessif doit être coréférent au sujet.

Intex et Unitex sont donc des outils très puissants et très performants en matière de traitement de corpus à la condition de disposer des ressources linguistiques nécessaires. En effet, la construction de grammaires locales pour décrire avec précision les séquences verbales figées permettrait de créer des transducteurs dont l'application pourrait aboutir à la reconnaissance automatique de telles séquences.

II/ Elaboration de l'application Verbalex

1. L'application Verbalex

1.1 Principes et Objectifs

La conception du logiciel Verbalex a pour objectif de réunir des outils pour le traitement de séquences figées et plus particulièrement des locutions verbales. Le logiciel vise donc l'extraction des séquences verbales présentant un caractère figé. La méthodologie adoptée pour la réalisation de ce programme diffère totalement de celle adoptée par les logiciels Intex et Unitex. Ces outils utilisent en effet des ressources déjà existantes pour produire une analyse lexicale et syntaxique après avoir identifié les éléments du texte. Ces ressources, comme nous l'avons vu, se présentent sous la forme de dictionnaires électroniques auxquels l'utilisateur n'a pas accès dans la mesure où ils sont destinés à permettre les traitements proposés par le logiciel. La méthodologie adoptée pour la construction du programme Verbalex se rapproche davantage de celle employée par Béatrice Daille pour le fonctionnement de l'outil ACABIT qui vise l'extraction de terminologie. La seule ressource utilisée pour procéder au traitement d'un corpus est ce corpus lui-même ainsi que les différentes informations linguistiques apportées par une opération d'étiquetage et de lemmatisation. La construction d'un dictionnaire électronique est le but visé par ce programme. Verbalex devrait donc permettre à terme de constituer un dictionnaire électronique des locutions verbales apparaissant dans un corpus ; ce dictionnaire serait ensuite accessible et modifiable par l'utilisateur. La construction de ce programme s'inscrit donc dans une conception totalement différente de celle adoptée par le LADL.

Verbalex prend donc en entrée un texte au format brut (.txt) et non balisé : l'insertion de clé ou tout marquage spécifique ne sont pas nécessaires, de même que l'encodage du fichier ne prend pas d'importance. Ce texte doit ensuite être traité afin de procéder à un filtrage des formes verbales. Ces formes verbales, une fois extraites, constituent une liste dont les composants ou items doivent être validés manuellement par l'utilisateur. La validation d'une séquence donnée conduit à la création d'une entrée dans un dictionnaire électronique. L'entrée du dictionnaire ainsi créée est ensuite intégrée à la structure du dictionnaire déjà existante. En effet, un dictionnaire électronique est construit à partir des résultats d'extraction produits par l'application. La structure de ce dictionnaire est prédéfinie. Ce dictionnaire électronique est disponible dans l'application indépendamment du ou des corpus traités. Le traitement des corpus ouverts en entrée permet l'enrichissement du dictionnaire.

L'architecture du logiciel pourrait être représentée de la manière suivante :

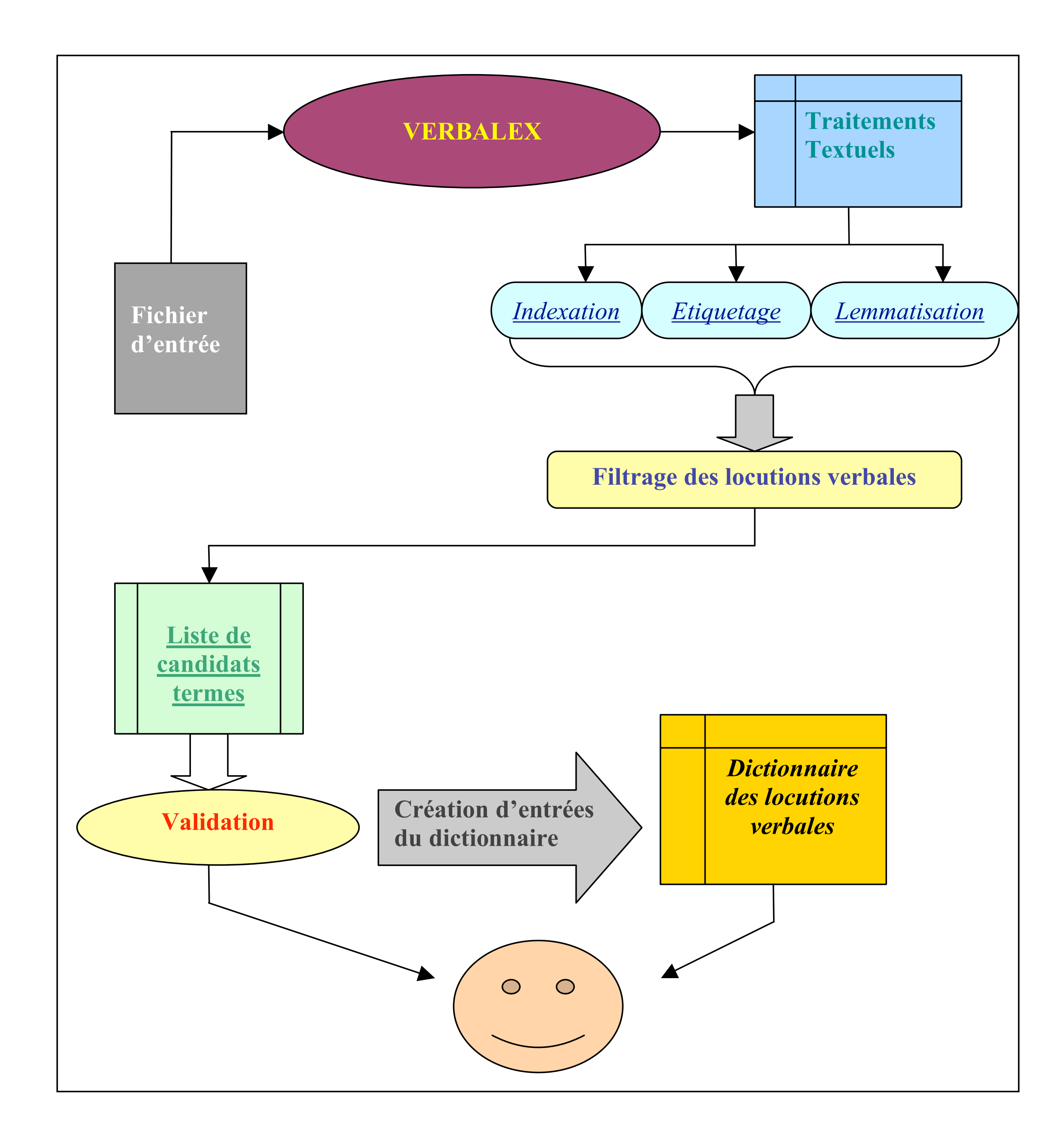


Figure 1.1.a) : Architecture du logiciel Verbalex

Nous pouvons voir au travers de ce schéma que l'étape de filtrage est précédée par une étape de prétraitement du texte qui doit donc être indexé, étiqueté et lemmatisé. Nous décrirons dans la section suivante en quoi consistent concrètement ces étapes. L'état du corpus résultant de ces diverses opérations sert de base à l'opération de filtrage. Cette opération va produire une liste de candidats termes qui seront ou non validés afin de constituer une entrée du dictionnaire. Cette entrée figurera ensuite dans le dictionnaire électronique existant déjà dans l'application et auquel l'utilisateur pourra apporter les modifications souhaitées.

1.2 Construction du programme

Langage

Le programme Verbalex a été écrit avec le langage de programmation Perl acronyme de « Practical Extraction and Report Language » que l'on pourrait traduire dans les grandes lignes par « Langage Pratique d'Extraction et d'Edition ».

Un langage de programmation diffère en de nombreux points d'une langue naturelle. En effet, le nombre d'unités composant une langue naturelle est potentiellement infini. Un langage de programmation est fini et limité même si des bibliothèques de scripts ou packages viennent l'agrémenter et le compléter. Les variations que connaissent un mot dans un langage de programmation sont limitées voir inexistantes. A chaque mot d'un langage de programmation est associé ou associable une et une seule catégorie, étiquette qui correspondent à un identifieur, c'est-à-dire un index, une variable, un entier,…

Le langage de programmation PERL a été créé par Larry Wall en 1986 pour gérer un système de « News » entre deux réseaux. PERL est un langage interprété qui n'est pas compilé et qui est donc moins rapide qu'un programme compilé. L'interpréteur PERL est nécessaire pour exécuter le programme. Ce langage est traditionnellement utilisé pour gérer des fichiers au format html notamment en ce qui concerne les scripts CGI. Le module Tk aussi dénommée Tool kit (Tk), proposé par PERL permet de créer et de gérer des interfaces graphiques. Perl/Tk utilise les caractéristiques orientées objet de PERL et n'est pas uniquement destiné aux utilisateurs de PERL mais convient aussi programmes écrits en langage C, Ada ou Python. Le langage Perl/Tk manipule des Widgets qui constituent des briques de base. L'écriture d'un programme en Perl/Tk consiste donc dans la création, la manipulation et le placement des widgets.

Description de l'interface

L'application « Verbalex » a dans un premier temps pris la forme d'un éditeur de texte classique du type Bloc-Notes ou Notepad, à savoir une zone de texte encadrée par un menu Fichier et un menu Edition proposant leurs fonctionnalités classiques respectives. Cette interface graphique a ensuite été enrichie de divers autres menus, mais également d'icônes et d'onglets. Des procédures permettant d'effectuer des recherches ou des remplacements ont tout d'abord figuré sous le menu Edition, puis ont été déplacées sous un menu spécifique intitulé « Traitements Textuels ». Ce menu comme son nom l'indique permet de procéder à spécifiques sur le fichier ouvert en entrée. Le fichier ouvert en entrée figure sous l'onglet Document.

Les menus Fichier, Edition et Traitements Textuels se présentent donc de la manière suivante :



Figure 1.2.a) : Verbalex : Les menus Fichier, Edition, et Traitements Textuels.

Les icônes qui apparaissent sur cette figure renvoient à des fonctionnalités disponibles dans les menus.

Les menus « Fichier » et « Edition » permettent respectivement de gérer l'ouverture, la fermeture et la sauvegarde de fichiers figurant sous l'onglet Document, et des fonctionnalités classiques telles qu'Annuler, ou les fameux copier-coller.



Figure 1.2.b.i) : Verbalex : Le menu Traitements Textuels.

Le menu « Traitements Textuels est subdivisé en deux sous-parties. La première partie comporte des opérations de recherche et de remplacement disponibles pour le Document. La seconde partie permet de prétraiter le texte pour procéder au filtrage des locutions verbales. Ce menu permet donc de procéder à la recherche d'un motif donné ou à des remplacements.

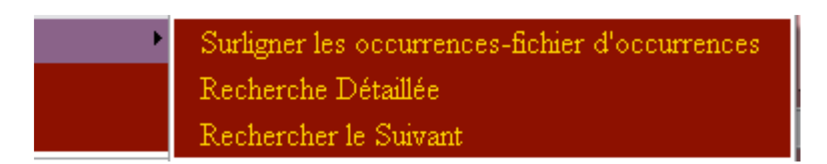


Figure 1.2.b.ii) : Verbalex : Menu en cascade de l'item « Rechercher ».

Les procédures de recherche et de remplacement sont inclues dans la bibliothèque de scripts Tk::TextUndo disponible avec le module Tk. Il n'a donc pas été nécessaire de créer des procédures manuellement : les procédures proposées par ce package permettent de

rechercher un élément donné occurrence par occurrence comme c'est le cas dans le Bloc-Notes, il en est de même pour les remplacements. Une procédure de recherche a toutefois été créée manuellement afin de surligner simultanément dans le document toutes les occurrences de la chaîne recherchée ; les lignes où apparaissent la forme recherchée sont aussi recensées dans une fenêtre de type popup qui s'ouvre quand la recherche est terminée et qui permet à l'utilisateur de sauvegarder le résultat obtenu. Les trois items composant la seconde partie de ce menu permettent d'accéder à l'onglet « Traitements Textuels » et aux sous-onglets correspondants. La figure suivante représente ces trois sous-onglets.

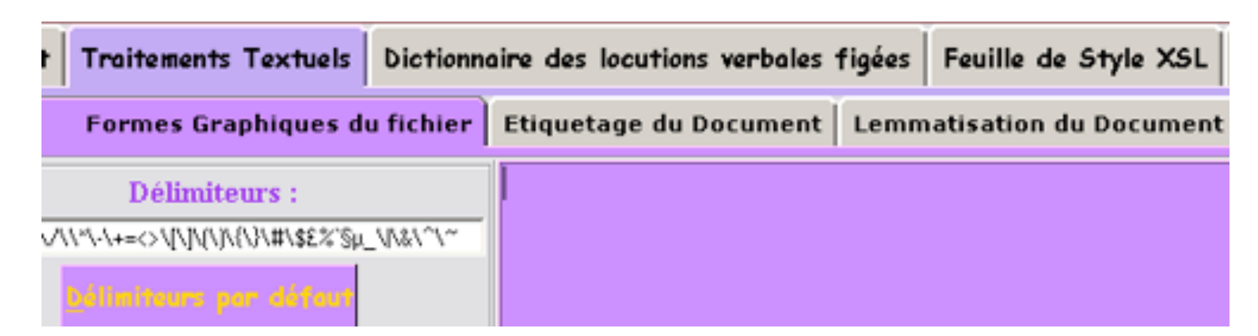


Figure 1.2.c) : Verbalex : Les sous-onglets de l'onglet Traitements Textuels.

Il est donc possible par l'intermédiaire de ce menu de préparer le corpus pour le filtrage des locutions verbales. Les différents sous-onglets permettent donc de prétraiter le texte en fournissant des informations sur ce texte par la construction d'un dictionnaire de formes graphiques, en procédant à l'étiquetage et à la lemmatisation du Document. Le « Dictionnaire de formes graphiques » qui apparaît dans l'onglet « Formes graphiques du fichier » recensent toutes les formes qui apparaissent dans le texte et de trier ces formes selon l'ordre alphabétique ou selon leur fréquence d'apparition. Tous les mots qui correspondent à ces formes et qui apparaissent dans le document sont donc répertoriés dans ce dictionnaire. Au terme « mot » correspond une définition purement typographique, à savoir une chaîne de caractères enclavée par deux blancs (ou espaces) comme nous l'avons vu. L'utilisateur peut choisir lui-même quel type de délimiteur il souhaite utiliser pour la constitution de ce dictionnaire. Le sous-onglet « Dictionnaire de formes graphiques » indique par ailleurs le nombre de lignes et de mots du fichier. Les sous-onglets « Etiquetage » et « Lemmatisation » permettent respectivement d'étiqueter et de lemmatiser le Document et d'afficher les fichiers obtenus dans les zones textuelles correspondantes. Ces sous-onglets sont placés dans cet ordre précis car l'étiquetage est une étape primordiale qui doit donc précéder l'étape de lemmatisation. Toutefois si la lemmatisation est lancée avant l'étiquetage, celui-ci est réalisé en arrière-plan (sans passer par le sous-onglet « Etiquetage ») pour procéder ensuite à la lemmatisation.



Figure 1.2.d) : Verbalex : Le menu Dictionnaire.

Le menu qui apparaît ensuite est le menu « Dictionnaire » aussi subdivisé en deux parties. Le premier item permet de lancer le filtrage des locutions verbales. La seconde partie de ce menu porte essentiellement sur le dictionnaire des locutions verbales lui-même et offre donc la possibilité de charger le dictionnaire, de rechercher une forme dans le dictionnaire, d'ajouter une entrée et de charger sa feuille de style. Le dernier item permet d'ouvrir la feuille de style du dictionnaire qui est accessible dans l'onglet « Feuille de Style ».



Figure 1.2.e) : Verbalex : Le menu Outils.

Le menu « Outils » offre des fonctionnalités supplémentaires concernant des fichiers au format html, XML ou XSL. Le dernier menu permet d'obtenir des informations sur l'application et d'ouvrir une fenêtre comportant l'aide pour le fonctionnement du logiciel.

Les différentes icônes apparaissant sous les menus renvoient aux fonctionnalités proposées par les items des menus en permettant donc un accès plus rapide.



Figure 1.2.f.i) : Verbalex : Les icônes.

Les icônes représentées sur cette figure permettent d'accéder plus rapidement aux fonctionnalités proposées par Verbalex.

	⇨	Accéder à l'onglet Traitements Textuels et au sous-onglet Lemmatisation.
	⇨	Lancer le filtrage des locutions verbales, une fois qu'un fichier a été chargé.
	⇨	Charger le dictionnaire des locutions verbales et accéder à l'onglet correspondant.
	⇨	Ajouter manuellement une entrée au dictionnaire.
	⇨	Charger la feuille de style du dictionnaire et accéder à l'onglet correspondant.

Figure 1.2.f.ii) : Verbalex : Fonctionnalités de ces icônes.

Les onglets et les sous-onglets permettent d'afficher le texte ainsi que les versions étiquetées et lemmatisées du texte, le dictionnaire électronique des locutions verbales (dans sa totalité ou par lettre) ainsi que la feuille de style du dictionnaire. Le dernier onglet permet de voir l'arborescence des fichiers sur l'ordinateur sur lequel l'utilisateur travaille. Outre l'affichage de ces divers fichiers, des options de recherche et d'enregistrement, ainsi que d'autres fonctionnalités sont disponibles dans ces onglets et ces sous-onglets. Des images écrans de ces différents onglets et de l'interface graphique de Verbalex sont présentées en Annexes.

Cette présentation de l'interface graphique de Verbalex utilisant des onglets et des sous-onglets s'avère judicieuse pour l'accès par l'utilisateur aux différents états du fichier et autres ressources. De plus, le fichier ouvert en entrée qui se présente donc sous l'onglet Document n'est pas altéré et ne subit donc aucune modification lors de son traitement (étiquetage, lemmatisation, …).

L'étiquetage

Comme nous l'avons vu dans la section consacrée à la présentation d'outils d'acquisition de terminologie, l'étiquetage constitue une étape importante dans le traitement automatique de corpus. Il s'agit d'une opération de base qui vise à étiqueter les formes pertinentes d'un texte ayant le statut d'unités de base. Cette étape permet de remédier à de nombreuses difficultés. En effet, les informations linguistiques ne sont pas déductibles de leur forme. Le –s en français, par exemple, ne permet pas toujours prédire une forme plurielle. Le nom « glas » ne correspond pas à une forme plurielle. Les ambiguïtés lexicales constituent une autre difficulté dans la mesure où la polysémie d'une forme implique l'association de plusieurs étiquettes. A une même forme peut correspondre deux catégories. La forme « porte » renvoie aussi bien à un verbe qu'à un nom pour désigner respectivement l'action de

« porter » et l'objet « porte » comme une « pore d'entrée ». La principale conséquence que peut avoir le processus d'étiquetage morphosyntaxique est la désambiguïsation – dans le logiciel Intex, cette étape d'étiquetage morphosyntaxique se fait à l'aide du dictionnaire électronique de désambiguïsation et présente le résultat produit sous forme de graphes.

La résolution des problèmes d'ambiguïtés suppose l'attribution de plusieurs étiquettes morphosyntaxiques pour les formes présentant plusieurs catégories. L'exemple classique utilisé pour illustrer ces problèmes d'ambiguïtés est le suivant :

Ex 1.a) : La petite brise la glace.

Selon l'étiquetage morphosyntaxique appliqué produira deux interprétations différentes :

La$_{[Det]}$ petite$_{[Adj]}$ brise$_{[N]}$ la$_{[Pron]}$ glace$_{[V]}$.

La$_{[Det]}$ petite$_{[N]}$ brise$_{[V]}$ la$_{[Det]}$ glace$_{[N]}$.

L'étiquetage morphosyntaxique consiste donc dans l'affectation automatique d'étiquettes morphosyntaxiques en fonction du contexte. L'étiqueteur utilisé dans cette application est l'étiqueteur TreeTagger[1].

TreeTagger est un étiqueteur probabiliste qui se distingue des étiqueteurs basant leur fonctionnement sur des règles. TreeTagger, contrairement aux étiqueteurs qui ont recours au modèle des chaînes de Markov cachées pour calculer les étiquettes des formes, procède à la construction d'un arbre de décision pour étiqueter un fichier.

TreeTagger prend en entrée un fichier de texte brut qui ne comporte aucun balisage ou marquage. Ce fichier doit néanmoins être formaté de manière à un présenter un mot par ligne pour que TreeTagger puisse procéder à l'étiquetage. Des procédures, plus précisément des scripts Perl, ont donc été créées pour réaliser ce formatage. Ces scripts ont été écrits de manière à conserver les signes de ponctuation. TreeTagger permet aussi de procéder à la lemmatisation du fichier après l'avoir étiqueté.

La lemmatisation

La lemmatisation consiste à remplacer une forme fléchie par son lemme. Le lemme constitue la forme de base d'un mot donné. La lemmatisation présente donc cette forme de base sans aucune marque de flexion (pluriel, désinence ou forme conjuguée d'un verbe). La lemmatisation permet de remplacer une forme actualisée par sa forme canonique. Le lemme ou la forme canonique d'un mot constitue une entrée de dictionnaire.

Le lemmatiseur Flemm[2] a été intégré à l'application Verbalex pour procéder à la lemmatisation du texte. Flemm est un programme écrit en Perl. Ce lemmatiseur prend en entrée un fichier au préalable étiqueté morphosyntaxiquement. Flemm lemmatise un fichier étiqueté par les étiqueteurs Brill ou TreeTagger. Ce lemmatiseur procède à une vérification

---

[1] TreeTagger : www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[2] Flemm : http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm

des couples mot-étiquette. Flemm corrige si besoin les étiquettes pour calculer le lemme du mot. Ce calcul se fait à base d'une centaine de règles. Ce système utilise un lexique réduit qui comporte différentes listes d'exceptions qui comportent elles-mêmes près de 3000 mots. L'utilisation de ce lexique permet à ce lemmatiseur de procéder à une analyse flexionnelle du mot lemmatisé sans avoir à prendre en compte le contexte de celui-ci.

TreeTagger ayant été paramétré pour produire l'étiquetage mais aussi la lemmatisation du fichier, Flemm vérifie aussi bien l'étiquette morphosyntaxique produite que le lemme qui lui est associé et corrige au besoin ces informations.

Les fichiers résultant de ces étapes d'étiquetage et de lemmatisation ont été retravaillés par l'intermédiaire de scripts Perl afin d'offrir un meilleur affichage. En effet, le fichier étiqueté et le fichier lemmatisé produits se présentaient comme le fichier pris en entrée, à savoir avec un mot par ligne : la forme était suivie d'une tabulation de son étiquette morphosyntaxique, d'une autre tabulation et de son lemme. L'écriture de ces scripts intégrés à Verbalex a donc consisté à remplacer ces tabulations par des underscores (« _ ») pour rendre les résultats plus lisibles.

Le filtrage des locutions verbales : mode d'emploi

L'utilisateur doit dans un premier temps charger un corpus en cliquant sur la deuxième icône ou en passant par le menu Fichier. Le corpus apparaît alors sous l'onglet Document. Une zone d'information située juste en dessous des icônes indique que le fichier a été ouvert.

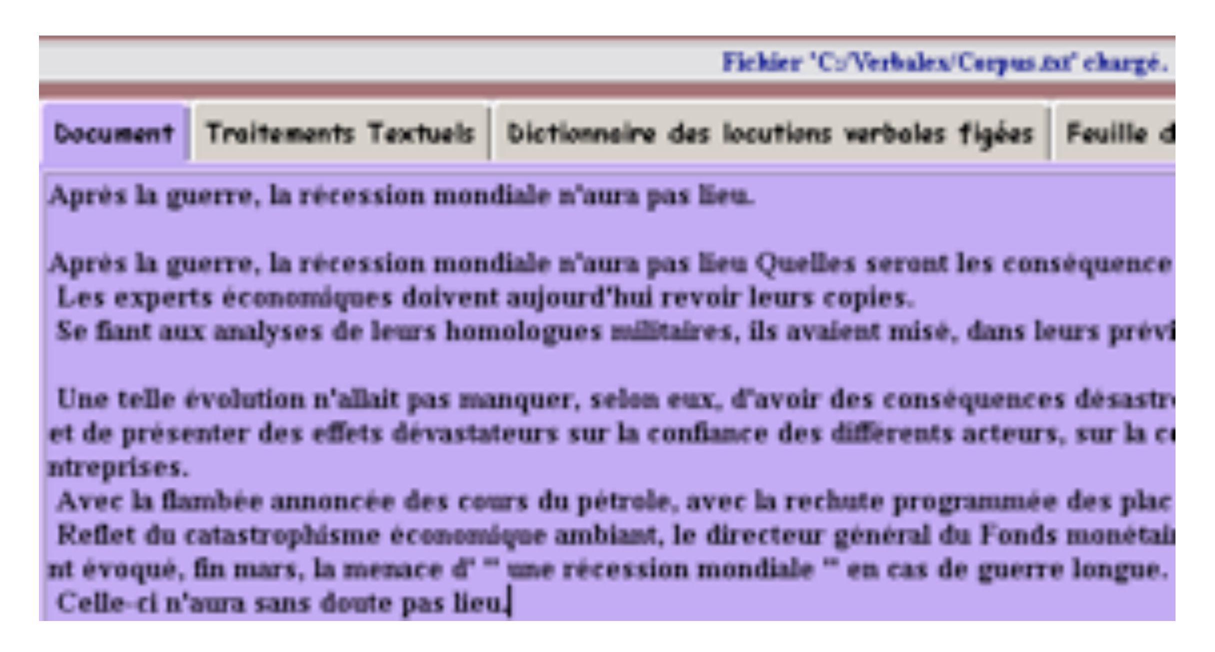


Figure 1.2.g) : Verbalex : Chargement du fichier.

Ce texte doit ensuite être étiqueté. Le fait de cliquer l'item « Etiquetage » du menu Traitements Textuels permettra de lancer l'étiquetage et de mettre en relief le sous-onglet correspondant.



Figure 1.2.h) : Verbalex : Etiquetage du fichier.

Le mode d'affichage choisi par défaut est l'affichage « texte en continu ».

Le fichier ainsi obtenu doit ensuite être lemmatisé en passant par le menu Traitements Textuels ou par le sous-onglet. Comme nous l'avons vu dans la description de l'interface il est possible de lancer la lemmatisation sans avoir à lancer l'étiquetage. Cette étape peut cependant s'avérer utile pour s'assurer que les étiquettes attribuées sont correctes.



Figure 1.2.i) : Verbalex : Lemmatisation du fichier.

Une fois que la lemmatisation est terminée, il faut s'assurer que l'affichage de la version lemmatisée est bien en « texte continu » avant de lancer le filtrage.

L'opération de filtrage des locutions verbales est lancée dans l'application via le menu « Dictionnaire » ou l'icône correspondante.



Figure 1.2.j) : Verbalex : Icône pour lancer le filtrage.

Ce filtrage est basé sur la définition de patrons syntaxiques. Ces différents patrons syntaxiques représentent les diverses structures morphologiques que peuvent présenter les locutions verbales : le tableau ci-dessous représente deux de ces patrons syntaxiques.

Patron Syntaxique	*Exemple*
Verbe Déterminant (indéfini) Nom **V Det N1**	*« Prendre la tangente »,* *« Prendre le large »*
Verbe Nom **V N1**	*« Avoir froid »,* *« Avoir lieu »*

Figure 1.2.k) : Verbalex : Patrons Syntaxiques de locutions verbales pris en compte par Verbalex.

Ces patrons syntaxiques, d'après l'étude que nous avons faite dans la première partie, représentent les structures morphosyntaxiques courantes de locutions verbales. Il existe bien entendu des structures beaucoup plus complexes telle que « **V Prep Det(pluriel) N1(pluriel)** » qui correspondrait à la locution « envoyer sur les roses ». Nous avons choisi de procéder uniquement au filtrage des patrons syntaxiques décrits dans la figure 1.2.a).

Ces deux patrons syntaxiques ont donc été formalisés sous forme de grammaire locale à base d'expression régulière. Cette méthode de filtrage s'apparente donc à celle adoptée par Daille pour la construction du programme ACABIT dans sa démarche d'acquisition de terminologie portant sur les adjectifs relationnels (voir la section 1.2.3.2 ACABIT). Nous avons donc tenté de décrire ces différents patrons syntaxiques par deux expressions régulières distinctes afin de prendre en compte l'insertion éventuelle de modifieurs.

Ces patrons peuvent correspondre à de nombreuses phrases dont le verbe peut aussi bien être libre, figé, ou support. Une fois que le filtrage est terminé, le résultat produit apparaît dans une fenêtre de type popup. Cette fenêtre comporte une liste de candidats termes, autrement dit une liste de locutions verbales potentielles.



Figure 1.2.l) : Verbalex : Liste de candidats termes produite par Verbalex

Les items de cette liste doivent donc être validés manuellement et individuellement pour s'assurer qu'il s'agit bien d'une locution verbale. L'item doit donc être sélectionné et l'utilisateur doit cliquer sur le bouton « Valider » pour lancer la validation d'une locution. Le bouton « Enregistrer » permet de sauvegarder la liste de candidats termes correspondant au Document pour la valider plus tard. La procédure de validation d'un item de cette liste de candidats termes entraîne donc l'apparition d'une grille de validation. La valeur de l'item sélectionné est récupérée pour figurer dans la grille de validation.



Figure 1.2.m) : Verbalex : Grille de validation.

Comme nous pouvons le voir sur cette figure, cette grille de validation reprend les critères et les formalisations que nous avons proposés dans l'analyse de corpus faite en 5.2 (Analyse de corpus). Le fait de cocher le trait « Opacité Sémantique » et au moins trois traits formels implique que la présence d'une locution verbale est fort possible. Le fait de cocher ces différents traits revient à dire que la suite correspondant à l'item de la liste est sémantiquement opaque et/ou que les différentes transformations indiquées sont impossibles pour cette suite. Dans le cas où ces critères ne seraient pas remplis, un message apparaît dans une boite de dialogue en indiquant que la séquence sélectionnée n'est pas une locution verbale.

Si la présence d'une locution est présupposée, la validation de la grille entraînera l'apparition d'une boite de dialogue permettant de créer une nouvelle entrée lexicale dans le dictionnaire de l'application. La séquence apparaîtra automatiquement dans la zone vedette de la fenêtre de création de l'entrée.

Nous verrons dans la section suivante consacrée au dictionnaire des locutions verbales de Verbalex comment se présente cette boite de dialogue.

1.3 Le dictionnaire des locutions verbales

Présentation

Tout dictionnaire comporte différentes entrées auxquelles correspondent des articles de dictionnaire. Ce projet de conception d'un outil ayant pour objet les locutions verbales, les entrées du dictionnaire correspondent à des séquences verbales dont le premier élément sera un verbe.

Ce dictionnaire des locutions verbales est donc construit par le traitement d'un corpus mais peut également être enrichi indépendamment de tout traitement de corpus. Ce dictionnaire est accessible en dehors de l'application dans la mesure où il s'agit d'un fichier au format « xml ».

XML (eXtensible Markup Language) est une version simplifiée de SGML, un autre langage à structure balisante. En effet, nous avons choisi ce format de fichier afin de pouvoir structurer et hiérarchiser au mieux les informations décrivant les différentes entrées du dictionnaire. Le choix de ce format permet notamment d'accéder à des fonctionnalités spécifiques à ce langage. Le standard XPATH s'est notamment révélé utile pour permettre un affichage du dictionnaire par lettre, de sorte que toutes les entrées correspondant à une même initiale apparaissent dans le même sous-onglet correspondant à cette initiale. XPATH est un langage de requêtes pour les documents XML. XPATH pourrait également permettre de filtrer les entrées en fonction des informations décrites dans leur article : l'utilisation de ce langage est donc parfaitement adaptée dans le cadre de ce travail dans la mesure où il offre à l'utilisateur la possibilité de définir des requêtes sur la forme des entrées ou sur les différents éléments de cette entrée. Il serait donc possible par exemple possible de filtrer toutes les définitions des entrées ou tous leurs lemmes.

Ce dictionnaire étant donc un document XML, une DTD (Definition Type of Document) a donc été créée pour définir la structure du document. Cette DTD constitue une grammaire du document et stipule quel type de contenu peut avoir un élément ou un attribut dans ce document. Le dictionnaire doit donc pour être valide avoir une structure conforme à celle définie par cette grammaire de document. Le cadre gauche de l'onglet Dictionnaire propose deux icônes auxquelles sont associées deux procédures différentes permettant respectivement de vérifier que la syntaxe du document est correcte et de s'assurer que le document est conforme à sa DTD, à savoir s'il est valide. La manière dont sont écrites les balises et la manière dont elles s'enchaînent sont définies par une syntaxe propre au langage XML. Ces deux icônes sont les suivantes :

	Vérifier la bonne formation du Document	⇨	Bibliothèque de Script XML::Parser
	Vérifier la validité du Document	⇨	Programme externe Rxp[3]

Le programme Rxp qui selon les options indiquées en argument permet de vérifier ou la bonne formation ou la validité du document a été intégré au programme Verbalex afin de valider le Document par rapport à sa DTD.

La DTD du dictionnaire construit par Verbalex se présente de la manière suivante :

```
<!-- DTD du document Dictionnaire.xml!-->

<!ELEMENT dictionnaire (entete,corps)>
<!ELEMENT entete (titre+,application+)>
<!ELEMENT corps (lettre+)>
<!ELEMENT lettre (initiale+,entree*)>

<!ELEMENT entree (lemme+,structure?,definition?,distribution?,remarque?)>

<!ELEMENT lemme (#PCDATA)>
<!ELEMENT structure (#PCDATA)>
<!ELEMENT definition (#PCDATA)>
<!ELEMENT distribution (#PCDATA)>
<!ELEMENT remarque (#PCDATA)>

<!ATTLIST lettre value CDATA #REQUIRED>
<!ATTLIST entree initiale CDATA #REQUIRED>
<!ATTLIST entree verbe CDATA #REQUIRED>

<!ELEMENT titre (#PCDATA)>
<!ELEMENT application (#PCDATA)>

<!ELEMENT initiale (#PCDATA)>
```

Figure 1.3.a) : Verbalex : DTD du Dictionnaire

Le fichier « Dictionnaire.xml » est donc constitué par un élément racine <dictionnaire>, cet élément a deux éléments fils <entete> et <corps>. L'élément <entete> fournit des informations sur le dictionnaire telles que le titre du document ou le nom de l'application. L'élément <corps> comprend vingt-six éléments qui correspondent aux lettres de l'alphabet. Les éléments <lettre> peuvent contenir plusieurs entrées ou ne pas en contenir.

La DTD permet donc de définir quelles seront les informations qui constitueront l'article de dictionnaire, à savoir quels sont les éléments et les contenus que reçoivent les entrées de ce dictionnaire. Nous allons voir dans la section suivante quelles sont ces informations.

---

[3] Rxp : http://www.cogsci.ed.ac.uk/~richard/rxp.html

L'entrée du dictionnaire

A une entrée du dictionnaire correspond un article de dictionnaire. Les informations contenues dans cet article sont réparties dans diverses zones. La constitution de ces zones a été inspirée par la structure d'un article de DEC *(voir la figure 2.2.b dans la partie B/, 2.2 Le Dictionnaire Explicatif et Combinatoire)*. Les dix zones définies par celui-ci n'ont pas été reprises dans leur intégralité, seules quelques-unes de ces zones ont donc été retenues pour constituer un article dans ce dictionnaire des locutions verbales.



Figure 1.3.b) : Verbalex : Entrée du dictionnaire pour la locution « casser sa pipe »

Cette figure représente la boite de dialogue proposée par Verbalex pour la création d'une nouvelle entrée de dictionnaire. Nous pouvons voir que cinq zones principales apparaissent. Ces cinq champs, « Vedette », « Structure Morphosyntaxique », « Définitions / Acceptions », « Propriétés Distributionnelles », et « Remarques / Nota Bene », correspondent respectivement à la zone vedette, la zone morphologique, la zone sémantique et à la zone Nota Bene de l'article-type d'un DEC. Les champs correspondant à ces zones ont été remplis sur cette figure en fonction des informations caractérisant la locution verbale « casser sa

pipe ». La validation de cette entrée va permettre l'écriture de ces informations dans le dictionnaire. Le résultat produit sera le suivant :

```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
  <!DOCTYPE dictionnaire (View Source for full doctype...)>
- <dictionnaire>
  - <entete>
      <titre>Dictionnaire des Locutions Verbales figées du français</titre>
      <application>VERBALEX</application>
    </entete>
  - <corps>
    - <lettre value="A">
        <initiale>A</initiale>
      </lettre>
      <!-- xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx! -->
    - <lettre value="B">
        <initiale>B</initiale>
      </lettre>
      <!-- xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx! -->
    - <lettre value="C">
        <initiale>C</initiale>
      - <entree initiale="C" verbe="Casser">
          <lemme>Casser sa pipe</lemme>
          <structure>V Poss(N0) N1</structure>
          <definition>Mourir</definition>
          <distribution>Coréférence du Possessif avec le GN Sujet</distribution>
          <remarque>Locution synonyme: "Passer l'arme à gauche"</remarque>
        </entree>
      </lettre>
      <!-- xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx! -->
```

Figure 1.3.c) : Entrée du dictionnaire pour la locution « casser sa pipe »

L'utilisation du langage XML permet donc de structurer ces informations en fonction de leurs natures.

Compte tenu de ces différentes informations, la structure arborescente de ce dictionnaire pourrait donc être représentée ainsi :

```
dictionnaire
├──entete
│  ├──titre Dictionnaire des Locutions Verbales figÃ©es du franÃ§ais
│  └──application VERBALEX
└──corps
   ├──lettre
   │  ├──lettre-value A
   │  └──initiale A
   ├──lettre
   │  ├──lettre-value B
   │  └──initiale B
   ├──lettre
   │  ├──lettre-value C
   │  ├──initiale C
   │  └──entree
   │     ├──entree-verbe Casser
   │     ├──entree-initiale C
   │     ├──lemme Casser sa pipe
   │     ├──structure V Poss(N0) N1
   │     ├──definition Mourir
   │     ├──distribution CorÃ©fÃ©rence du Possessif avec le GN Suj
   │     └──remarque Locution synonyme: "Passer l'arme Ã  gauche"
   ├──lettre
   │  ├──lettre-value D
   │  └──initiale D
   ├──lettre
   │  ├──lettre-value E
   │  └──initiale E
   ├──lettre
   │  ├──lettre-value F
   │  └──initiale F
   ├──lettre
   │  ├──lettre-value G
   │  └──initiale G
```

Figure 1.3.d) : Structure arborescente du Dictionnaire

La feuille de style du dictionnaire

Une feuille de style a été associée à ce dictionnaire afin de mettre en forme ce dictionnaire des locutions verbales. La feuille de style a été créée afin d'offrir une meilleure lisibilité des informations fournies par le dictionnaire. Le document transformé par la feuille sera plus accessible qu'un document XML.

Cette feuille de style correspond à un fichier au format « .xsl ». L'association de cette feuille de style au dictionnaire permet donc de transformer ce document au format « .xml » en page Web au format « .html ». Cette feuille de style a donc été écrite avec le langage XSL (eXtensible Style Language) qui est un langage de formatage et de transformation.

Le document xsl a donc été construit en fonction des informations contenues dans le dictionnaire et en fonction de la manière dont elles s'organisent. Cette feuille de style établit

donc des règles qui s'appliquent à des nœuds bien précis de l'arborescence du dictionnaire pour définir des modèles. Ces modèles définissent les éléments de style qui caractérisent les nœuds auxquels ils s'appliquent.



Figure 1.3.e) : Format html du Dictionnaire de locutions verbales : l'entête du dictionnaire.

Des règles spécifiques ont été définies afin de formater l'entête du dictionnaire comme il est possible de le voir sur cette figure. Des ancres ont notamment été définies afin de créer des liens vers chaque lettre pour permettre un accès rapide aux données.

L'application de la feuille de style au dictionnaire des locutions verbales produit donc le résultat suivant au niveau de l'entrée du dictionnaire :

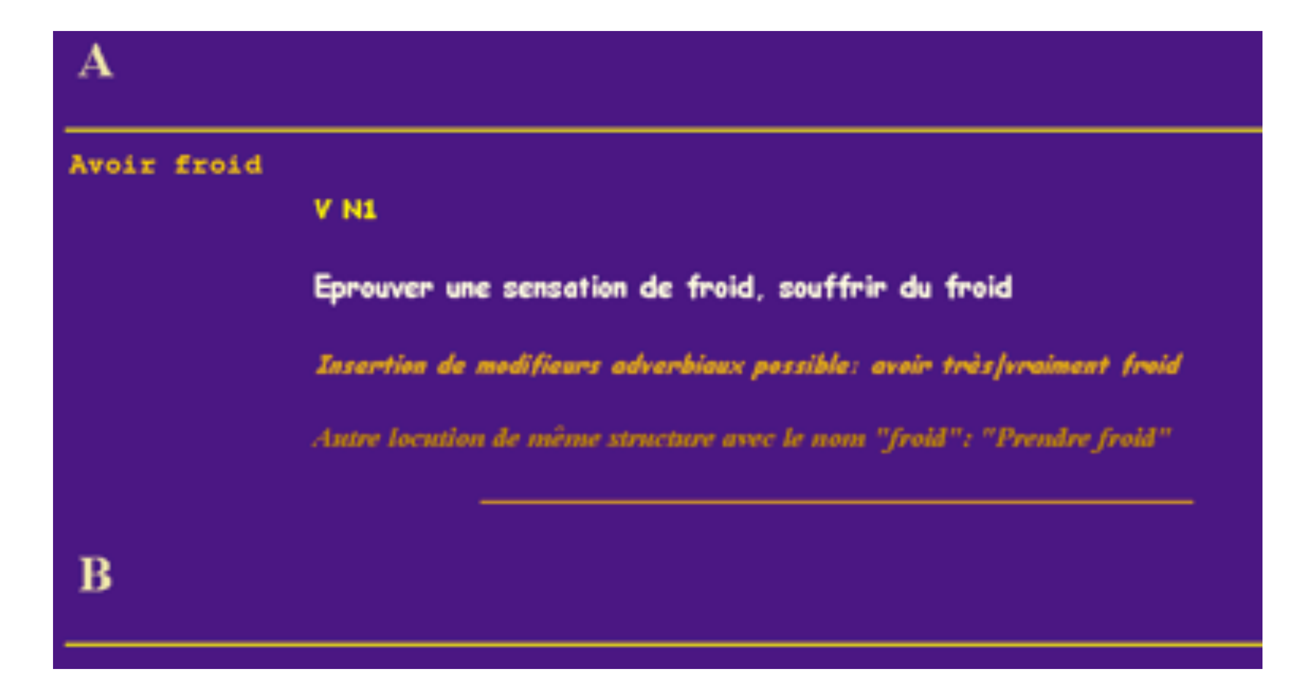


Figure 1.3.f) : Format html du Dictionnaire de locutions verbales : une entrée du dictionnaire.

Des polices et des styles différents ont été attribué aux diverses informations de l'entrée du dictionnaire afin de les distinguer au mieux.

2. Perspectives

L'élaboration de ce programme relevait davantage d'une étude expérimentale que de l'intention de trouver LA solution qui pourrait favoriser le traitement automatique des séquences figées, à fortiori celui des locutions verbales.

Les améliorations qui pourraient être apportées à ce programme sont donc nombreuses. La première amélioration, qui n'est pas la moindre, résiderait dans le perfectionnement de la procédure de filtrage. Les expressions régulières utilisées pour décrire les patrons syntaxiques potentiels des locutions doivent être retravaillés pour affiner davantage le filtrage et parer aux phénomènes de discontinuité. Le corpus choisi dans le cadre de ce travail étant d'une taille réduite, l'application de méthodes statistiques pour effectuer un tri préférentiel dans la liste de candidats termes n'était plus de mise. L'utilisation de telles méthodes, bien qu'efficaces uniquement sur des corpus de taille importante, pourrait apporter une amélioration dans cette démarche de filtrage.

Le programme Verbalex tel qu'il est actuellement construit permet uniquement un accès en écriture du dictionnaire dans le processus d'extraction des locutions verbales. En effet, une procédure de consultation dans le dictionnaire devrait être ajoutée. Le fait de vérifier que les items de la liste de termes candidats ne figurent pas déjà dans le dictionnaire permettrait d'affiner quelque peu le filtrage et de réduire la liste des candidats termes. De même si une expression figure plusieurs fois dans le corpus, une procédure de tri dans la liste de candidats termes réduirait le nombre de ces derniers.

La structure de l'entrée du dictionnaire pourrait éventuellement être enrichie par davantage d'informations linguistiques telles que l'ajout d'une zone phonologique par exemple donnant des informations sur la prononciation ou la prosodie propre à l'entrée du dictionnaire.

Les entrées du dictionnaire sont regroupées dans celui-ci en tant qu'éléments fils de l'élément « lettre » dont la valeur correspond à la première lettre du verbe de la locution verbale. Les différentes entrées commençant par la même lettre ne sont pas triées par ordre alphabétique, étant donné que le programme a été écrit de telle sorte que la dernière entrée saisie et validée figure juste sous l'initiale. Une procédure de tri par ordre alphabétique pourrait donc être utile pour organiser et structurer au mieux ce dictionnaire.

Enfin nous avons tenté de procéder ici à un traitement des locutions verbales dont le verbe peut être figé ou seulement figurer dans une expression figée (dont le groupe nominal objet serait figé). Selon la position que nous avons adoptée, les verbes qui entrent dans un groupement verbe-complément(s) figé sans pour autant être eux-mêmes figés sont considérés comme entrant dans des locutions verbales. Les verbes qui entrent dans ces locutions peuvent aussi bien être libre, figé ou support. Le champ d'application de ce programme pourrait donc être élargi au traitement de verbes et de constructions aussi courantes que les constructions à verbe support qui présentent une structure très ressemblante à celle des locutions verbales.

Conclusion

Nous avons tenté de décrire tout au long de ce travail quelles étaient les propriétés linguistiques des séquences figées et plus particulièrement celles des locutions verbales. Nous avons vu au travers de cette étude que les théories et les termes pour désigner le figement étaient très divergents et qu'elles ont évolué au cours du temps. Les travaux actuels et notamment ceux menés par le LADL ont fortement contribué à une meilleure description de ce phénomène propre aux langues naturelles.

Ces mêmes travaux du LADL ont également permis de faire évoluer la situation au niveau du traitement automatique. En effet, la démarche descriptive exhaustive de toutes les formes du français, entreprise par le LADL, a permis la conception d'outils exploitant les ressources produites par leur étude. Intex est l'un de ces outils. Bien que dans le cadre de cette étude, cet outil ne se soit pas révélé fort efficace, Intex pourrait produire des résultats satisfaisants. La condition pour parvenir à ces résultats serait donc de construire des ressources dont des transducteurs permettant de décrire chacune des séquences figées et pour ce faire, d'exploiter au mieux les fonctionnalités proposées par ce logiciel.

Le traitement automatique des expressions figées s'avère donc être une tâche rigoureuse et difficile. La création de l'application Verbalex nous a notamment permis de prendre conscience de cette difficulté. Au-delà de la difficulté due au traitement des locutions verbales, nous avons également pu constater que la création d'un outil quel que soit son but ou son champ d'application doit être mûrement réfléchie afin d'offrir à l'utilisateur l'interface la plus cohérente, accessible et efficace possible. En effet, quels que soient la méthodologie ou l'outil utilisé les connaissances et les aptitudes linguistiques de l'utilisateur sont mises à contribution pour procéder à un traitement automatique de séquences figées.

La conclusion qui s'impose donc à l'issue de ce travail est que ce type de démarche ne peut se passer de la présence et de la participation d'un linguiste. Si les travaux et les recherches menés actuellement tendent à automatiser le plus possible le traitement des séquences figées, un traitement entièrement automatisé, à court terme, n'est pas envisageable.

Bibliographie

OUVRAGES

- BERNARD Georges, " Les locutions verbales françaises", La Linguistique, 1974, Vol 10-2, pages 5-17.
- BOURIGAULT Didier, "Analyse syntaxique locale pour le repérage de termes complexes dans un texte", TAL, 1993, Volume 34, N°2, pages105-117.
- BRUN Carole, JACQUEMIN Christian, SEGOND Frédérique, "Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale", TAL, 2001, Volume 42 "Lexiques Sémantiques", n3, pages 667-690.
- COURTOIS Blandine, Un système de dictionnaires électroniques pour les mots simples du français, Langue Française, 1990, N°87, pages 11-22.
- CURAT Hervé, La Locution verbale en français moderne : essai d'explication psycho-systématique. Québec, P.U. Laval, 1986, (P. Vachon-L'Heureux).
- DAILLE Béatrice, "L'identification en corpus d'adjectifs relationnels: une piste pour l'extraction automatique de terminologie", TAL, 2001, Volume 42 "Lexiques Sémantiques", n3, pages 815-832.
- DANLOS Laurence, « La morphosyntaxe des expressions figées », Langages, 1981, n°63.
- DARMESTETER Arsène, « Traité de la formation des mots composés », Paris, Bouillon, 1874.
- DUGAS André, La création lexicale et les dictionnaires électroniques. Langue Française, 1990, N°87, pp. 23-329.
- GAATONE David, « Les locutions verbales : pour quoi faire ? », Revue Romane, 1981, volume 16, Copenhague,
- GAATONE David, "Les locutions verbales et les deux passifs du français", Langages, 1993, N°109, pages 37-52.

- GIRY-SCHNEIDER, Jacqueline. Les prédicats nominaux en français : les phrases simples à verbe support, Genève : Droz, 1987, 391 p.

- GROSS Gaston, « Les expressions figées en français, noms composés et autres locutions », Paris, Ophrys,1996,.

- GROSS Gaston, " Lexicographie et Grammaire", Cahiers de lexicologie, 1981, Vol 39-2, pages 35-46.

- GROSS Maurice, « Les limites de la phrase figée », Langages, 1988, n°90, p7-22, Larousse, Paris.

- GROSS Maurice, "Sur les déterminants dans les expressions figées", Langages, 1985, Vol 79, Larousse, Paris.

- GUILLET Alain, "Reconnaissance des formes verbales avec un dictionnaire minimal", Langue Française, 1990, Vol 87, Larousse, Paris.

- HABERT Benoît, JACQUEMIN Christian, "Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques", 1993, Volume 34 « Traitements automatiques de la composition nominale » (TAL), N°2, pages 5-41.

- HABERT B., NAZARENKO A. and SALEM A., « Les Linguistiques de corpus », Armand Colin, 1997.

- HEID, Ulrich; FREIBOTT, Gerhard, " Collocations dans une base de données terminologique et lexicale ", *Meta*, 36, 1, mars 1991, p. 77-91.

- LECLERE Christian, "Organisation du Lexique-Grammaire des verbes français", Langue Française, 1990, N°87, Paris, Larousse.

- LEHMAN Alise et MARTIN-BERTHET Françoise, « Introduction à la lexicologie, sémantique et morphologie », 2000, Paris, Collection Lettres Sup., Nathan Université.

- LE PESANT D., MATHIEU-COLAS M., « Introduction aux classes d'objets », Langages, 1998, n°131, Larousse, Paris.

- LINDBERG Lars, Les Locutions verbales figées dans la langue française, thèse Upsal pour le doctorat par Lars Lindberg, 1898.

- MEL'CUK Igor, CLAS André, POLGUERE Alain, "Introduction à la lexicologie explicative et combinatoire", 1995, Louvain-la-Neuve: Editions Duculot (Coll. Universités Francophones)

- MEJRI Salah, « Le figement lexical, descriptions linguistiques et structuration sémantique », 1997, Publications de la Faculté des lettres de la Manouba.

- MISRI Georges, "Approches du figement linguistique: critères et tendances", La Linguistique, 1987, Vol 23, pages 72-85, Paris.

- PIERREL Jean-Marie, « Ingénierie des Langues », Éditions Hermes Science, 2000, Collection "Information - Commande - Communication", 360 pages, ISBM 2-7462-0113-5.

- REY Alain et CHANTREAU Sophie, « Dictionnaire d'Expressions et Locutions », Paris, collection « Les Usuels », 1989, Dictionnaires Le Robert.

- SILBERTZEIN Max, « Le dictionnaire électronique des mots composés », Langue Française, 1990, Vol 87, pages 71-83, Larousse, Paris.

LIENS INTERNET

- « Construire et accéder à une base de données d'expressions figées à partir des ressources de la toile », Gaël Dias, Ludovina Carapinha, Rosa Trinidade, Susana Mota, Marco Ribeiro et Jorge Dias, Université de la Beira Interior (Portugal) :
  http://www.di.ubi.pt/~ddg/publications/tia2003.pdf

- Méthodologie pour la création d'un dictionnaire distributionnel dans une perspective d'étiquetage lexical semi-automatique, Delphine Reymond, équipe DELIC, Université de Provence :
  http://loria.fr/projets/TALN/actes/Recital/pleniere/reymond.pdf

- La description des collocations et leur traitement dans les dictionnaires, Marleen Laurens
  http://www.Kuleuven.ac.be/vlr/994colloc.htm

- Les expressions idiomatiques : de la marginalité à la reconnaissance, Claudia Maria Xatara, Université de l'Etat de Sao Paulo, Brésil :
  http://fdlm.org/fle/article/319/idiomatique.php3

- L’interrogation de bases de données comme application des classes d’objets, Béatrice Bouchou, Julien Lerat, Denis Maurel, LI Université François Rabelais :
  http://li.univ-tours.fr/taln-recital-2001/Actes/tome1_PDF/partie2_p30_322/art09_p113_122.pdf

- Défigements sémantiques en contexte, François Rastier, CNRS :
  http://www.revue-texto.net/Inedits/Rastier_Defigements.html_0

- La cooccurrence en T.A.L : Dis-moi qui tu fréquentes et je te dirai qui tu es, DelphineReymond, DELIC, Université de Provence :
  http://www.up.univ-mrs.fr/wpsycle/ColloqueEcriture/docinformatique/reymond.html

- Localisation et analyse d’expressions figées dans de grands corpus, Pierre Dupont, Cédrick Fairon (CENTAL) :
  http://www.info.ucl.ac.be/enseignement/memoires/2003-2004/pdupont_tal1.html

- Accéder au sens culturel par le décontextualisation : le cas des énoncés médiatiques, Teta Simeonido-Christido, Université de Thessalonique (Grèce) (colloque 10 et 11 Mars 1997), §22 La force discursive de la locution phraséologique :
  http://crim.inalco.fr/recomu/colloque/07.phtml

- Un modèle HMM pour la détection des mots composés dans un corpus textuel, Lakhdar Remaki et Jean Guy Meunier, LANCI, Université du Québec à Montréal :
  http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/67/67.pdf

- Les N-grams de caractères pour l’aide à l’extraction de connaissances dans des bases de données textuelles multilingues, Ismail Biskri et Sylvain Delisle :
  http://www.li.univ-tours.fr/taln-recital-2001/Actes/tome1_PDF/partie2_p30_322/art07_p93_102.pdf

- Dictionnaires distributionnels et étiquetage lexical de corpus, Delphine Reymond, EquipeDELIC :
  http://www.li.univ-tours.fr/taln-recital-2001/Actes/tome1_PDF/partie4_p403_482/art8_p473_482.pdf

- Un étiquetage morphologique pour une résolution des ambiguïtés morphologique en anglais, Gaëlle Birocheau :
  http://www.sciences.univ-nantes.fr/irin/taln2003/articles/birocheau.pdf

- Unitex : traitement de corpus par dictionnaires électroniques et grammaires, Sébastien Paumier, IGM, Ecole d'été de Corpus, Caen, 15 juin 2004 : http://www.u-grenoble3.fr/lebarbe/elc/supports/paumier.pdf

- Probabilistic Part-of-speech Tagging Using Decision Trees , Helmut Schmid, IMS-CL: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

- Improvements In part-of-speech Tagging with an application to German, Helmut Schmid, IMS-CL: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf

- Locutions verbales pouvant être automatisées facilement : http://www.cs.lth.se/home/Pierre_Nugues/memoires/christophe/these/source/rul.loc

- Introduction à la lexicologie explicative et combinatoire, Mel'cuk, 1995, Claire Gardent: http://webloria.loria.fr/~gardent/teaching/semLex/melcuk4.pdf

- Recherches lexicographiques à l'OLST : Dictionnaire explicatif et combinatoire du français (DECFC), Alain Polguère, 1er Septembre 2000 : http://www.ling.umontreal.ca/olst/Fr/DECFC.html

- Informatisation du Dictionnaire Explicatif et Combiantoire, Gilles Sérasset (GETA-CLIP), Actes de TALN, 12-13 juin 1997: http://www-clips.imag.fr/geta/gilles.serasset/taln97-serasset.pdf

- Bibliothèques d'automates finis et grammaires context-free : de nouveaux traitements informatiques, Mathieu Constant, LADL, RECITAL 2001, 2-5 juillet 2001: http://www.li.univ-tours.fr/taln-recital-2001/Actes/tome1_PDF/partie4_p403_482/art3_p425_434.pdf

- Dictionnaire électronique des mots composés (DELAC), Sébastien Paumier : http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/delac.html

- Extraction terminologique avec Intex, Ibekwe-SanJuan Fidelia, URI-INIST, 4ème Journées Intex, Bordeaux, 11-12 juin 2001 : http://fidelia1.free.fr/intex01.pdf

- Représentation et utilisation de connaissances dans un système d'aide à l'apprentissage lexical, Thierry Selva et Fabrice Issac : http://lifc.univ-fcomte.fr/RECHERCHE/P7/pub/JCSC96/JCSC96.html

- Automates à états finis, S. Ratté : http://www.seg.etsmtl.ca/sylvie/LOG310/Cours/Theme02.pdf

- Grammaires régulières : http://www.univ-nancy2.fr/poincare/perso/rebuschi/cours/iup2/IUP-coursOC-2.6.pdf

- Automates finis et langage régulier : http://www.univ-nancy2.fr/poincare/perso/rebuschi/cours/iup2/IUP-coursOC-2.4.pdf

- Une bibliothèque d'opérateurs linguistiques pour la consultation de base de données en langue naturelle, Béatrice Bouchou et Denis Maurel, Conférence TALN 1999, Cargèse, 12-17 juillet 1999 : http://talana.linguist.jussieu.fr/taln99/ps/A5/A5.pdf

- Réflexions sur l'homographie et la désambiguïsation des formes les plus fréquentes, Anne Dister, JADT 2000, 5éme Journées Internationales d'Analyse Statistique des données textuelles : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/17/17.pdf

- Commentaires sur Unitex : http://laseldi.univ-fcomte.fr/intex/Unitex.htm

- Développer des grammaires Locales de levées d'ambiguïtés pour INTEX, Aurore Ferret, Séverine Gedzelmann : http://www.u-grenoble3.fr/idl/cursus/enseignants/tutin/Intex.htm

- Formation à Unitex : http://www-igm.univ-mlv.fr/~laporte/proj/TP2003UnitexD1.htm

- Grammaires Locales, Sébastien Paumier : http://www-igm.univ-mlv.fr/~paumier/DEA/Cours%209%20-%20Grammaires%20locales.pdf

- Enjeux linguistiques et informatiques des expressions figées : http://www.limsi.fr/Individu/habert/Publications/Fichiers/habert91b/BH_C1.html

- Analyse et filtrage : http://www.limsi.fr/Individu/habert/Publications/Fichiers/habert91b/BH_C2.html

OUTILS

Intex :

Logiciel :
http://intex.univ-fcomte.fr/downloads/

Manuel d'utilisation :
http://intex.univ-fcomte.fr/downloads/Manual.pdf

Unitex

Logiciel :
http://www-igm.univ-mlv.fr/~unitex/download.html

Manuel d'utilisation :
http://www-igm.univ-mlv.fr/%7Eunitex/manuelunitex.pdf

Flemm

http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm

TreeTagger

Logiciel :
www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Liste des Etiquettes Morphosyntaxiques:
http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html

Perl :

http://www.activestate.com/Products/ActivePerl/

Rxp :
http://www.cogsci.ed.ac.uk/~richard/rxp.html

Annexes

Glossaire

Automate à état fini :

Type particulier de transducteur à état fini qui ne produit pas d'information et qui ne considère que les informations produites en entrée.

Actualisation :

L'actualisation permet d'inscrire un prédicat dans son contexte. Un verbe par exemple est actualisé par sa conjugaison. Un nom est actualisé par l'emploi d'un verbe support.

Composition :

Procédé de formation de nouvelles unités lexicales généralement opposé à la dérivation, à partir d'éléments lexicaux ayant une existence autonome dans la langue.
Mot composé : mot contenant deux ou plus de deux morphèmes lexicaux.

Compositionnalité (Compositionnel) :

Une suite est dite compositionnelle si le sens de cette suite est déductible à partir des éléments composants reliés par une relation syntaxique spécifique.

Concordance :

La concordance d'une séquence est un index qui représente toutes les occurrences de cette séquence dans son contexte.

Défigement :

Ce phénomène désigne la perte du caractère figé d'une séquence donnée afin de créer un effet humoristique. Les défigements sont de deux sortes :

- Le défigement peut être syntaxique ou sémantique : la structure de la séquence figée est modifiée de telle sorte que les lectures figée et compositionnelle se superposent en créant une ambiguïté.
- Le défigement peut se produire par contexte : l'emploi d'une séquence figée qui apparaît dans un contexte inattendu crée un défigement dû à l'incidence du contexte sur cette séquence.

Délimiteurs :

Signes ou symboles permettant de délimiter des unités de base. Les espaces, les signes de ponctuation et les retours à la ligne sont par exemple des délimiteurs. Les délimiteurs utilisés par les programmes diffèrent selon leurs besoins.

Etiquetage:

Opération consistant à attribuer une catégorie morphosyntaxique (étiquette) à une forme donnée d'un corpus.

Figement :

- Le figement peut être syntaxique : les possibilités combinatoires ou transformationnelles normalement disponibles pour les constructions libres sont interdites pour les constructions figées.
- Le figement peut être sémantique : les sens de la suite est opaque, c'est-à-dire non compositionnel.

Il existe des degrés de figement.

Grammaire Locale :

Représentation par automate de structures linguistiques complexes qui ne sont pas formalisables dans le lexique-grammaire et les dictionnaires DELA. Ces grammaires locales sont visuellement présentées sous forme de graphes.

Langage de programmation :

Un langage de programmation est fini et limité et permet de compiler un programme informatique. A chaque mot d'un langage de programmation est associé ou associable une et une seule catégorie, étiquette qui correspondent à un identifieur, c'est-à-dire un index, une variable, un entier…

Lemmatisation:

Opération consistant à remplacer une forme fléchie par sa forme canonique ou son lemme.

Lexie :

- Unité de base de l'étude lexicologique s'inscrivant dans le cadre de la théorie Sens-Texte. Ces lexies sont décrites dans un dictionnaire qui correspond à la structure initiale d'un DEC (Dictionnaire Explicatif et Combinatoire).
- « Unité lexicale mémorisée » d'après la terminologie de Bernard Pottier (1962). Une lexie peut être simple, composée, ou complexe.

Locution :

Etymologiquement « manière de dire ». Séquences inférieures au niveau de la phrase présentant un caractère figé. Les auteurs parlent généralement de locutions adjectivales, verbales, adverbiales ou prépositives.

Mot :

- Suite composée d'un ou de plusieurs morphèmes et faisant sens.
- Chaîne de caractères comprise entre deux espaces.

Mot simple:

- Unité qui ne peut être décomposée en plusieurs morphèmes. S'oppose aux mots dérivés.
- "Une unité de texte définie sur l'alphabet des codes ASCII et ne comportant aucun séparateur (ni trait d'union, ni blanc ni apostrophe) ", d'après les travaux du LADL. S'oppose au mot composé.

N-Gram:

Un n-gram de caractères correspond à une suite de n caractères. Un bi-grams désigne une séquence de 2 caractères dans la mesure où n= 2.

Opacité :

Une suite est dite opaque quand le sens des éléments composants ne permet pas d'obtenir le sens global de la suite. Dire d'une séquence qu'elle est sémantiquement opaque équivaut à dire qu'elle est non-compositionnelle.

Phraséologie :

Discipline linguistique ayant pour objet les lexies complexes qui sont constituées de plusieurs mots graphiques et qui se comportent comme des lexies simples, qui sont traditionnellement appelées mots composés, locutions verbales, locutions adjectivales,… ou encore idiotisme.

Polylexicalité (Polylexical) :

Une suite est dite« polylexicale » quand elle est composée de plusieurs éléments lexicaux qui ne jouent pas de rôle extérieur à la séquence.

Synapsie :

Séquences de mots figées d'après la terminologie d'Emile Benveniste : « unité de signification composée de plusieurs morphèmes lexicaux ».

Synthème :

D'après la terminologie d'André Martinet (1967) : « unités linguistiques dont le comportement syntaxique est strictement identique à celui des monèmes avec lesquels ils commutent, mais qui peuvent être conçus comme formés d'éléments sémantiquement identifiables. »

Transducteur à état fini :

Graphe représentant un ensemble de séquences en entrée et leur associant des séquences produites en sortie. Un transducteur est un automate à état fini qui se distingue cependant de ce dernier dans la mesure où il comporte aussi bien une bande de lecture qu'une bande d'écriture qui permettent de fournir des informations sur une forme du texte. Un transducteur est constitué d'un ensemble de nœuds dont un nœud initial et un nœud terminal ; tous les autres nœuds représentent les formes du texte.

Terme :

Objet d'étude de la Terminologie au même titre que le morphème constitue l'objet d'étude de la morphologie. Un terme a pour fonction de représenter une notion, un concept dans un domaine de connaissance D.

Token:

Unités d'informations correspondant traditionnellement aux mots simples mais pouvant également correspondre à des n-grams de caractères.

Verbe Support:

Verbe de sens général qui n’a pas de fonction prédicative. Les verbes supports sont aussi dits verbes opérateurs. Les verbes supports apportent les informations de temps, de personne, de nombre et d’aspect à un prédicat nominal. La combinaison formée par le verbe support et le prédicat nominal avec lequel il est construit peut être paraphrasée par un verbe simple sémantiquement équivalent.

« Zone Fixe » :

Partie d'une expression figée qui admet un nombre de fixe mots simples, même si ces mots sont susceptibles de variations morphologiques.

Corpus

Extraits de l'édition électronique du journal « Le MONDE » du 13 avril 2003

Après la guerre, la récession mondiale n'aura pas lieu.

Après la guerre, la récession mondiale n'aura pas lieu. Quelles seront les conséquences pour la croissance mondiale de la guerre en Irak ?
Les experts économiques doivent aujourd'hui revoir leurs copies.
Se fiant aux analyses de leurs homologues militaires, ils avaient misé, dans leurs prévisions les plus récentes, sur un enlisement du conflit. Une telle évolution n'allait pas manquer, selon eux, d'avoir des conséquences désastreuses pour les économies des grands pays industrialisés et de présenter des effets dévastateurs sur la confiance des différents acteurs, sur la consommation des ménages et sur l'investissement des entreprises. Avec la flambée annoncée des cours du pétrole, avec la rechute programmée des places boursières, la grande dépression était pour demain. Reflet du catastrophisme économique ambiant, le directeur général du Fonds monétaire international (FMI), Horst Köhler, avait solennellement évoqué, fin mars, la menace d' " une récession mondiale " en cas de guerre longue. Celle-ci n'aura sans doute pas lieu. Mais si l'hypothèse militaire à l'origine de ce scénario économique noir a été invalidée par les faits, si trois semaines seulement ont suffi à la coalition américano-britannique pour mettre à terre le régime de Saddam Hussein, faut-il pour autant, en matière d'anticipation de croissance, passer d'un profond pessimisme à un optimisme béat ? UNE "CHANCE" D'ÉVITER LA CRISE. Les analystes sont d'accord entre eux pour prédire la remontée un peu partout, au cours des prochaines semaines, des indicateurs de confiance économique, qu'il s'agisse du moral des ménages ou du climat des affaires. Ce rebond, reflet du soulagement ressenti devant le dénouement rapide de la guerre en Irak, mettrait fin à une longue période de reflux ininterrompu qui avait fait descendre ces baromètres à des records de faiblesse depuis quatre ans en France, depuis dix ans aux Etats-Unis. Les économistes soulignent aussi les effets positifs qu'aura rapidement pour l'activité économique la baisse des prix énergétiques. Mais s'il y a consensus sur le court terme, apparaissent des désaccords sur le long terme. Pour les uns, la fin des incertitudes géopolitiques qui minaient l'économie mondiale depuis plus d'un an va permettre aux grands pays industrialisés de renouer durablement avec un rythme de croissance élevé. C'est l'avis exprimé par le Prix Nobel d'économie Milton Friedman, pour qui la guerre en Irak "va sans aucun doute plutôt stimuler la conjoncture". "Je ne vois pas pourquoi l'économie ne devrait pas reprendre de l'élan une fois les incertitudes liées au conflit irakien disparues", ajoute le père de la théorie monétariste. De façon plus cynique, la victoire rapide des troupes américaines et britanniques, en provoquant un électrochoc puissant, serait une chance unique pour les économies occidentales de ne pas connaître la crise que subit le Japon depuis treize ans et l'éclatement de sa bulle spéculative boursière et immobilière. Tout le monde n'est pas aussi optimiste. Pour beaucoup, la chute de Saddam Hussein ne règle rien aux problèmes de fond de l'économie mondiale : mauvaise santé financière des entreprises, finances publiques à la dérive, déficit des comptes extérieurs américains (503 milliards de

dollars en 2003), rigidité du marché du travail et pression fiscale trop élevée en Europe, fragilité bancaire au Japon et en Allemagne, instabilité du marché des changes, déprime persistante des Bourses, etc. Alors que la perspective de la guerre en Irak avait, depuis des mois, fait diversion et servi d'explication facile au ralentissement de la croissance, toutes ces faiblesses, maintenant que Bagdad est tombée, vont resurgir. On s'orienterait, du coup, vers des années de croissance lente, molle. C'est l'opinion exprimée mercredi 9 avril par l'économiste en chef du FMI, Kenneth Rogoff, pour qui"il est peu probable que la croissance hésitante enregistrée actuellement dans le monde se transforme brusquement en une vigoureuse reprise économique". S'il paraît délicat de mesurer quel sera l'impact réel du conflit irakien sur la croissance économique mondiale, du moins quelques enseignements d'ordre "géoéconomique" et "géomonétaire", c'est-à-dire l'équilibre des forces entre les différentes zones géographiques, semblent d'ores et déjà se dégager. L'avancée rapide de leurs soldats, les preuves données de leur supériorité technologique en matière d'armements et de télécommunications, l'efficacité de leur stratégie militaire ont probablement contribué à redonner aux Etats-Unis - du moins dans les milieux financiers, très sensibles aux symboles de puissance - une partie du prestige que l'explosion de la bulle boursière, le scandale Enron, l'explosion de leurs déficits et le ralentissement de leur économie leur avaient enlevé. Il n'est pas impossible que l'effondrement du régime Saddam Hussein aide à faire oublier celui du Nasdaq et permette à l'économie américaine de retrouver sa prééminence des années 1990. Si cette dernière n'est pas vraiment contestée dans les chiffres - le PIB américain a augmenté de 2,4 % en 2003, trois fois plus vite que dans la zone euro -, elle commençait à l'être dans les esprits. Autant un enlisement du conflit aurait mis à mal l'image d'hyperpuissance économique des Etats-Unis, autant une guerre courte risque de la renforcer. INFLUENCE ACCRUE DE LA MAISON BLANCHE La hausse du dollar qui a suivi la prise de Bagdad semble confirmer cette analyse. Le renforcement du billet vert devrait toutefois être de courte durée dans la mesure où Washington, désireux de stimuler son économie et de réduire le déficit de ses comptes extérieurs, ne souhaite pas une telle évolution. Là encore, la victoire américaine en Irak, en augmentant le pouvoir d'influence de la Maison Blanche sur les marchés, devrait aider cette dernière à guider le dollar vers les niveaux qui lui conviennent. D'autant que le conflit a, en parallèle, semé une zizanie sans précédent en Europe. Les dirigeants de la zone euro seraient aujourd'hui bien en peine de riposter aux Etats-Unis si ceux-ci décidaient, en guise de représailles monétaires contre la France et l'Allemagne, de déclencher une vraie guerre des taux de change. Pierre-Antoine Delhommais

Astro Boy et la passion des Nippons pour les humanoïdes.

Astro Boy et la passion des Nippons pour les humanoïdes Yokohama de notre correspondant Créé en 1951 par Osamu Tezuka, qui en fait d'abord le héros d'une bande dessinée, avant de porter à l'écran ce qui deviendra la première série animée de la télévision nippone en 1963, Astro Boy, ou Tetsuwan Atom en japonais, est l'enfant- robot le plus célèbre de l'archipel. Dans l'histoire originale, le professeur Tenma invente un robot à l'image de son fils mort et l'active le... 7 avril. Depuis longtemps attendue par les fans de la série, la journée du 7 a donné lieu à toutes sortes de célébrations, d'un défilé costumé devant la gare de Takadanoba, à Tokyo (site du laboratoire du professeur Tenma), à la "naissance" d'un Astro Boy grandeur nature à Robodex, pilotée par nul autre que Macoto Tezca, fils d'Osamu Tezuka, en passant par la diffusion d'anciennes et nouvelles séries par la télé nippone. En véhiculant l'image du robot gentil, Astro Boy s'apparente à une sorte de mythe fondateur de la robotique humanoïde

nipponne. "Tous ceux qui s'occupent de robots ici ont en tête Astro Boy : il n'y a pas de tabou au Japon comme il peut y en avoir en Occident sur le fait de recréer une forme humaine et de vouloir rivaliser avec Dieu", explique Junji Suzuki, de Mitsubishi. Abandonné par son créateur dans un cirque pour robots, Astro Boy sera recueilli par un autre scientifique, avant d'embrasser une carrière de super-héros au service de la paix et de l'harmonie entre les humains et les machines. Dans le Japon de l'ère préélectronique, l'image du robot au bon cœur répondait aux angoisses de la course à la modernisation. Elle recoupe aujourd'hui les préoccupations des chercheurs nippons en quête - littéralement - d'une robotique à visage humain. Un laboratoire de l'université des sciences de Tokyo étudie ainsi la restitution, par un visage artificiel, des expressions humaines. Et celui de l'université de Waseda travaille sur les émotions. Les robots humanoïdes font même l'objet d'un programme du METI, sur cinq ans, qui a débuté en 1998 et qui regroupe une douzaine de sociétés et autant d'universités. Celles-ci ont mis au point plusieurs robots capables de travailler dans des environnements dangereux pour l'homme ou de piloter des engins de construction en étant manipulés à distance. Ces robots, tel HRP-2, qui mesure 1, 54 m et pèse 58 kg, qui est capable de se relever et d'aider un être humain à transporter des objets, pourraient servir de plates-formes adaptables à diverses applications industrielles. Chez Sony, l'équipe de Toshitada Doi, le créateur d'Aibo, planche sur SDR-4XII, un nouveau robot de divertissement de forme humanoïde qui pourrait bientôt être commercialisé pour " le prix d'une voiture de luxe". SDR est un petit robot (58 cm pour 7 kg) qui fait office de compagnon de jeu. Il est capable de chanter et de danser, mais aussi de converser en puisant dans une base de données de 60 000 mots. Honda continue le développement de son robot bipède Asimo. Star nationale au Japon, Asimo serre la main des chefs d'Etat étrangers et participe à toutes sortes d'événements. Le but affiché du constructeur automobile est de familiariser le grand public avec les robots. En l'an un du robot domestique, Astro Boy n'est déjà plus seul au monde. B. Pe.

Analyse du corpus

		Critère sémantique	Critère formel (syntaxique)						
Séquence	<Lemme>	Opacité Sémantique	Passif	Extraction	Détachement	Pron.	Relativation	Interrogation	Nature de la séquence (LOC V V Sup V libre)
N'aura pas lieu	<avoir lieu>	+	-	-	-	-	-	-	Loc Vb
Seront	X	X	X	X	X	X	X	X	V libre
Doivent	X	X	X	X	X	X	X	X	V libre
Revoir leurs copies	<revoir Poss copie>	-	-	-	?	-	-	+	Loc Vb
Avaient misé sur un enlisement …	X	X	X	X	X	X	X	X	V libre
N'allait pas manquer d'avoir des conséquences	X	X	X	X	X	X	X	X	V libre
Présenter	X	X	X	X	X	X	X	X	V libre
Etait pour demain	X	X	X	X	X	X	X	X	V libre
Avait évoqué	X	X	X	X	X	X	X	X	V libre
a été invalidée	X	X	X	X	X	X	X	X	V libre
Ont suffi	X	X	X	X	X	X	X	X	V libre
Mettre à terre le régime de Saddham Hussein	<mettre à terre>	-/+	+	+	?	-	+	-	Loc Vb
Passer d'un …à …	X	X	X	X	X	X	X	X	V libre
Eviter la crise	X	X	X	X	X	X	X	X	V libre
Sont d'accord	<être d'accord>	-	-	-	+	+	-	-	Loc Vb
Prédire la remontée des indicateurs…	X	X	X	X	X	X	X	X	V libre
Mettrait fin à une période	<mettre fin à>	-	-	-	-	-	-	-	Loc Vb
Avait fait descendre le	X	X	X	X	X	X	X	X	V libre

baromètre …									
Soulignent	X	X	X	X	X	X	X	X	V libre
Aura	X	X	X	X	X	X	X	X	V libre
Il y a	X	X	X	X	X	X	X	X	V libre
Apparaissent	X	X	X	X	X	X	X	X	V libre
Minaient	X	X	X	X	X	X	X	X	V libre
Va permettre	X	X	X	X	X	X	X	X	V libre
Renouer avec …	X	X	X	X	X	X	X	X	V libre
Va stimuler	X	X	X	X	X	X	X	X	V libre
Je ne vois pas	X	X	X	X	X	X	X	X	V libre
Ne devrait pas	X	X	X	X	X	X	X	X	V libre
Reprendre de l'élan	<prendre de l'élan>	-	+	+	?	+	+	+	V Sup
Ajoute	X	X	X	X	X	X	X	X	V libre
Serait	X	X	X	X	X	X	X	X	V libre
Connaître la crise	<connaître la crise>	-	+	+	+	+	+	+	V Sup
Subit	X	X	X	X	X	X	X	X	V libre
N'est pas optimiste	X	X	X	X	X	X	X	X	V libre
Ne règle rien aux problèmes …	X	X	X	X	X	X	X	X	V libre
Avait fait diversion	<faire diversion>	-	+	+	-	?	+	?	V Sup
Avait servi	X	X	X	X	X	X	X	X	V libre
Est tombée	X	X	X	X	X	X	X	X	V libre
Vont resurgir	X	X	X	X	X	X	X	X	V libre
S'orienterait vers	X	X	X	X	X	X	X	X	V libre
Il est peu probable que	X	X	X	X	X	X	X	X	V libre
Se transforme en …	X	X	X	X	X	X	X	X	V libre
Il parait délicat de	X	X	X	X	X	X	X	X	V libre
Mesurer l'impact	<mesurer l'impact>	-	-	-	-	-	-	-	Loc Vb
Semblent se dégager	X	X	X	X	X	X	X	X	V libre
Ont contribué à	X	X	X	X	X	X	X	X	V libre
Redonner une partie du prestige	X	X	X	X	X	X	X	X	V libre
Avaient enlevé (une partie du prestige)	X	X	X	X	X	X	X	X	V libre
Il n'est pas impossible	X	X	X	X	X	X	X	X	V libre
Aide à	X	X	X	X	X	X	X	X	V libre

Faire oublier …	X	X	X	X	X	X	X	X	V libre
Permette	X	X	X	X	X	X	X	X	V libre
A augmenté de	X	X	X	X	X	X	X	X	V libre
Commençait à	X	X	X	X	X	X	X	X	V libre
Aurait mis à mal	<mettre à mal>	-	+	-	-	-	-	-	Loc Vb
Risque de	X	X	X	X	X	X	X	X	V libre
A suivi	X	X	X	X	X	X	X	X	V libre
Semble confirmer	X	X	X	X	X	X	X	X	V libre
Devrait être de courte durée	<être de courte durée>	-	-	-	?	-	-	-	Loc Vb
Stimuler son économie	X	X	X	X	X	X	X	X	V libre
Réduire le déficit	X	X	X	X	X	X	X	X	V libre
Ne souhaite pas une telle évolution	X	X	X	X	X	X	X	X	V libre
Devrait aider à guider …	X	X	X	X	X	X	X	X	V libre
Niveaux qui lui conviennent	X	X	X	X	X	X	X	X	V libre
A semé une zizanie sans précédent	<semer la zizanie>	-/+	+	+	+	?	+	-	V Sup
Seraient en peine de riposter	X	X	X	X	X	X	X	X	V libre
Décidaient de déclencher	X	X	X	X	X	X	X	X	V libre
En fait le héros	X	X	X	X	X	X	X	X	V libre
Porter à l’écran	<porter à l’écran>	-	+	+	-	-	-	-	V Sup
Deviendra	X	X	X	X	X	X	X	X	V libre
Est	X	X	X	X	X	X	X	X	V libre
Invente	X	X	X	X	X	X	X	X	V libre
L’active	X	X	X	X	X	X	X	X	V libre
A donné lieu à	<donner lieu à >	-	-	-	-	-	-	-	Loc Vb
S’apparente à	X	X	X	X	X	X	X	X	V libre
S’occupent de	X	X	X	X	X	X	X	X	V libre
Il n’y a pas	X	X	X	X	X	X	X	X	V libre
Il peut y (en avoir)	X	X	X	X	X	X	X	X	V libre
Recréer une forme humaine	X	X	X	X	X	X	X	X	V libre
Vouloir rivaliser avec	X	X	X	X	X	X	X	X	V libre
Explique	X	X	X	X	X	X	X	X	V libre
Sera recueilli	X	X	X	X	X	X	X	X	V libre

Embrasser une carrière de super héros	<embrasser une carrière de N2>	+	-	+	-	-	-	-	Loc Vb
Répondait aux angoisses	X	X	X	X	X	X	X	X	V libre
Recoupe les préoccupations	X	X	X	X	X	X	X	X	V libre
Etudie la restitution	X	X	X	X	X	X	X	X	V libre
Travaille sur les émotions	X	X	X	X	X	X	X	X	V libre
Font l'objet d'un programme	<faire l'objet de >	-	-	-	-	-	-	-	Loc Vb
A débuté	X	X	X	X	X	X	X	X	V libre
Regroupe une douzaine de sociétés	X	X	X	X	X	X	X	X	V libre
Mis au point des robots	<mettre au point>	-/+	+	-	-	-	+	-	Loc Vb
Travailler	X	X	X	X	X	X	X	X	V libre
Piloter de engins	X	X	X	X	X	X	X	X	V libre
Mesure 1,54 m	X	X	X	X	X	X	X	X	V libre
Est capable de se relever	X	X	X	X	X	X	X	X	V libre
Aider un humain à transporter	X	X	X	X	X	X	X	X	V libre
Pourraient servir de plateforme	X	X	X	X	X	X	X	X	V libre
Planche sur un nouveau robot	X	X	X	X	X	X	X	X	V libre
Pourrait être commercialisé	X	X	X	X	X	X	X	X	V libre
Est un petit robot	X	X	X	X	X	X	X	X	V libre
Fait office de compagnon de jeu	<faire office de N2>	-	-	-	-	-	-	-	Loc vb
Est capable de chanter	X	X	X	X	X	X	X	X	V libre
Continue	X	X	X	X	X	X	X	X	V libre
Serre la main des chefs d'etats	X	X	X	X	X	X	X	X	V libre
Participe à	X	X	X	X	X	X	X	X	V libre
Familiariser le public avec les robots	X	X	X	X	X	X	X	X	V libre

N'est plus seul au monde	X	X	X	X	X	X	X	X	V libre

Liste des expressions figées du corpus produite par Intex

{ Abandonné, abandonner la partie,V+C1d}
{ Abandonné, abandonner la partie.V+C1d}
allait {ne-pas,pas.ADV+NEG}
aura {ne-pas,pas.ADV+NEG}
{ but, boire le coup,V+C1d}
{ cours, courir la gueuse,V+C1d} Avec la flambée annoncée des
{ cours, courir la gueuse,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir la gueuse,V+C1d} Avec la flambée annoncée des
{ cours, courir la gueuse,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir la prétentaine,V+C1d} Avec la flambée annoncée des
{ cours, courir la prétentaine,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir la prétentaine,V+C1d} Avec la flambée annoncée des
{ cours, courir la prétentaine,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir le cotillon,V+C1d} Avec la flambée annoncée des
{ cours, courir le cotillon,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir le cotillon,V+C1d} Avec la flambée annoncée des
{ cours, courir le cotillon,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir le guilledou,V+C1d} Avec la flambée annoncée des
{ cours, courir le guilledou,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir le guilledou,V+C1d} Avec la flambée annoncée des
{ cours, courir le guilledou,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les filles,V+C1d} Avec la flambée annoncée des
{ cours, courir les filles,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les filles,V+C1d} Avec la flambée annoncée des
{ cours, courir les filles,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les garçons,V+C1d} Avec la flambée annoncée des
{ cours, courir les garçons,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les garçons,V+C1d} Avec la flambée annoncée des
{ cours, courir les garçons,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les honneurs,V+C1d} Avec la flambée annoncée des
{ cours, courir les honneurs,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les honneurs,V+C1d} Avec la flambée annoncée des
{ cours, courir les honneurs,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les jupons,V+C1d} Avec la flambée annoncée des
{ cours, courir les jupons,V+C1d} {à,.PREP} {le,.DET:ms}
{ cours, courir les jupons,V+C1d} Avec la flambée annoncée des
{ cours, courir les jupons,V+C1d} {à,.PREP} {le,.DET:ms}
{ court, courir la gueuse,V+C1d} sur le
{ court, courir la gueuse,V+C1d} sur le
{ court, courir la prétentaine,V+C1d} sur le
{ court, courir la prétentaine,V+C1d} sur le

{ court, courir le cotillon,V+C1d} sur le
{ court, courir le cotillon,V+C1d} sur le
{ court, courir le guilledou,V+C1d} sur le
{ court, courir le guilledou,V+C1d} sur le
{ court, courir les filles,V+C1d} sur le
{ court, courir les filles,V+C1d} sur le
{ court, courir les garçons,V+C1d} sur le
{ court, courir les garçons,V+C1d} sur le
{ court, courir les honneurs,V+C1d} sur le
{ court, courir les honneurs,V+C1d} sur le
{ court, courir les jupons,V+C1d} sur le
{ court, courir les jupons,V+C1d} sur le
devrait {ne-pas,pas.ADV+NEG}
est {ne-pas,pas.ADV+NEG}
{ forces, forcer la mesure,V+C1d}
{ forces, forcer la note,V+C1d}
règle {rien,.PRO+NEG}
{ remontée, remonter la pente,V+C1d} de ÉVITER LA CRISE Les analystes sont de accord entre eux pour prédire la
{ remontée, remonter la pente,V+C1d} de ÉVITER LA CRISE Les analystes sont de accord entre eux pour prédire la
{ remontée, remonter la pente,V+C1d} de ÉVITER LA CRISE Les analystes sont de accord entre eux pour prédire la
{ reprendre, reprendre le dessus,V+C1d}
{ reprise, reprendre le dessus,V+C1d} actuellement dans le monde se transforme brusquement en une vigoureuse
{ sont remontée, remonter la pente,V+C1d} de ÉVITER LA CRISE Les analystes de accord entre eux pour prédire la
souhaite {ne-pas,pas.ADV+NEG}
vois {ne-pas,pas.ADV+NEG}
y a {ne-pas,pas.ADV+NEG}

Liste des mots composés du corpus produite par Unitex

à distance,.A+EPC+z1

à distance,.ADV+PC+z1

à l'économie,.A+EPDETC+z1

à l'écran,.A+EPDETC+z1

à l'image de,.PREP+EPCDN+z1

à l'origine de,.PREP+EPCDN+z1

à l'origine,.ADV+PDETC+z1

à la dérive,.A+EPDETC+z1

à mal,.A+EPC+z1

à terre,.A+EPC+z1

à visage humain,.A+EPCA+z1

activité économique,.N+NA+z1:fs

alors que,alors.CONJS+4

alors que,alors.CONJS+5

années de,année de.NDET+Dnom13

au cours,.ADV+PCDN2+z1

aujourd'hui,.ADV+z1

autant que,.CONJS+8

autant que,autant.CONJS+4

autant que,autant.CONJS+8

avant de,avant.PREP+Prépconjs+5

baisse des prix,.N+NDN+z3:fs

baisse des prix,.N+NDN:fs

bande dessinée,.N+NA+Conc+z1:fs

base de données,.N+NDN+Conc+z3:fs

billet vert,.N+NA+Conc+z1:ms

cas de guerre,.N+NDN+z3:ms:mp

cas de guerre,.N+NDN:ms:mp

celle-ci,celui-ci.PRON+Dém+z1:fs

celles-ci,celui-ci.PRON+Dém+z1:fp

ceux-ci,celui-ci.PRON+Dém+z1:mp

consommation des ménages,.N+NDN:fs

constructeur automobile,.N+NA+Hum+z1:ms

cours des,.NDET+Dnom10

cours des,cour de.NDET+Dnom7

cours des,cours de.NDET+Dnom10

cours du,cour de.NDET+Dnom7

court terme,.N+AN+z3:ms

court terme,.N+AN:ms

croissance économique,.N+NA+z1:fs

d'abord,.ADV+PC+z1

dans l'histoire,.ADV+PCDN+z1

dans la mesure,.ADV+PCDN+z1

dans le monde,.ADV+PDETC+z1

de confiance,.ADV+PC+z1

de courte durée,.A+EPAC+z1

de la maison,.A+EPDETC+z1

de là,.ADV+PC+z1

défilé costumé,.N+NA+z2:ms

depuis longtemps,.ADV+PC+z1

directeur général,.N+NA+Hum+z1:ms

douzaine de,.NDET+Dnom1

douzaine de,.NDET+Dnom10

du moins,.ADV+PDETC+z1

économie américaine,.N+NA+z1:fs

économie mondiale,.N+NA+z1:fs

économies occidentales,économie occidentale.N+NA+z1:fp

effets dévastateurs,effet dévastateur.N+NA+E01+z1:mp

en cas de,.PREP+PCDN+z1

en cas de,.PREP+PCDN1+z1

en cas de,en cas.PREP+Prépconjs+6

en cas,.ADV+Advconjs+6

en cas,.ADV+PCDN+z1

en fait d',en fait de.PREP+PCDN+z1

en fait d',en fait de.PREP+PCDN1+z1

en fait,.ADV+PC+z1

en fait,.ADV+PCDN+z1

en guise de,.PREP+EPCDN+z1

en guise de,.PREP+PCDN+z1

en guise de,.PREP+PCDN1+z1

en guise,.ADV+PCDN+z1

en matière,.ADV+PCDN+z1

en Occident,.A+EPC+z1

en parallèle,.ADV+PC+z1

en passant par,.ADV+PV+z1

en passant,.ADV+PV+z1

en peine de,.PREP+EPCPQ+z1

engins de construction,engin de construction.N+NDN+Conc+z3:mp

entre eux,.A+EPC+z1

équipe de,.NDET+Dnom10

Etats-Unis,.N+NA+HumColl+z3:mp

Etats-Unis,.N+PR+Top+Ppays+IsoUs:mp

Etats-Unis,Etats-Unis d'Amérique.N+Loc:mp

être humain,.N+NA+Hum+z1:ms

experts économiques,expert économique.N+NA+Hum+z2:mp

finances publiques,.N+NA+z1:fp

fonds monétaire,.N+NA+z1:ms

forme humaine,.N+NA+z1:fs

gare de,.NDET+Dnom7

grand public,.N+AN+HumColl+z3:ms

grand public,.N+AN+HumColl:ms

grandeur nature,.ADV+PCA+z1

grandeur nature,.N+NN+z1:fs

grands pays industrialisés,grand pays industrialisé.N+ANA+Conc:mp

indicateurs de,indicateur de.NDET+Dnom9

là encore,.ADV+PCA+z1

la journée,.ADV+PDETC+z1

la main,.GN+A1+z1

la paix,.GN+A1+z1

la télé,.GN+A1+z1

la télévision,.GN+A1+z1

laboratoire de,.NDET+Dnom7

laboratoire du,laboratoire de.NDET+Dnom7

le long,.ADV+PCDN+z1

le plus,.ADV+PCDN+z1

le plus,.ADV+PCDN2+z1

long terme,.N+AN+z3:ms

long terme,.N+AN:ms

main des,main de.NDET+Dnom10

main des,main de.NDET+Dnom7

marché des changes,.N+NDN:ms

marché des,marché de.NDET+Dnom7

marché du travail,.N+NDN+z3:ms

marché du travail,.N+NDN:ms

marché du,marché de.NDET+Dnom7

mauvaise santé financière,.N+ANA+z3:fs

milieux financiers,.N+NA+HumColl+z1:mp

milliards de,.NDET+Dnom6

milliards de,milliard de.NDET+Dnom1

milliards de,milliard de.NDET+Dnom6

optimisme béat,.N+NA+z1:ms

partie du,partie de.NDET+Dnom12b

pas de,.NDET+Dnom2

pays industrialisés,pays industrialisé.N+NA+Conc+HumColl+z1:mp

période de,.NDET+Dnom13

places boursières,place boursière.N+NA+Conc+z1:fp

plates-formes,plate-forme.N+AN+Conc+z3:fp

plates-formes,plate-forme.N+AN+Conc:fp

pour beaucoup,.ADV+PC+z1

pour demain,.A+EPC+z1

pression fiscale,.N+NA+z3:fs

prix Nobel,.N+NN+Hum+z1:ms:mp

problèmes de fond,problème de fond.N+NDN:mp

ralentissement de la croissance,.N+NDN:ms

régime de,.NDET+Dnom10

reprise économique,.N+NA+z1:fs

robot domestique,.N+NA+Conc+z2:ms

rythme de croissance,.N+NDN+z3:ms

rythme de croissance,.N+NDN:ms

s'il y a,.ADV+PF+z1

Saddam Hussein,.N+Hum+NPropre:ms

sans aucun doute,.ADV+PAC+z1

sans doute,.ADV+PC+z1

sans précédent,.A+EPC+z1

santé financière,.N+NA+z1:fs

service de,.NDET+Dnom10

stratégie militaire,.N+NA+z1:fs

super-héros,.N+AN+Hum+z3:ms:mp

super-héros,.N+AN+Hum:ms:mp

taux de change,.N+NDN:ms:mp

un peu partout,.ADV+PAC+z1

un peu,.ADV+PDETC+z1

une fois,.ADV+PDETC+z1

université de,.NDET+Dnom7

université des,université de.NDET+Dnom7

voiture de luxe,.N+NDN+Conc+z3:fs

voiture de luxe,.N+NDN+Conc:fs

voiture de,.NDET+Dnom7

zones géographiques,zone géographique.N+NA+Conc+z1:fp

Codes utilisés par les dictionnaires électroniques DELA

Code	Signification	Exemples
A	adjectif	fabuleux
ADV	adverbe	réellement, à la longue
CONJC	conjonction de coordination	mais
CONJS	conjonction de subordination	puisque, à moins que
DET	déterminant	ses, trente-six
INTJ	interjection	adieu, mille millions de mille sabords
N	nom	prairie, vie sociale
PREP	préposition	sans, à la lumière de
PRO	pronom	tu, elle-même
V	verbe	continuer, copier-coller

Codes grammaticaux

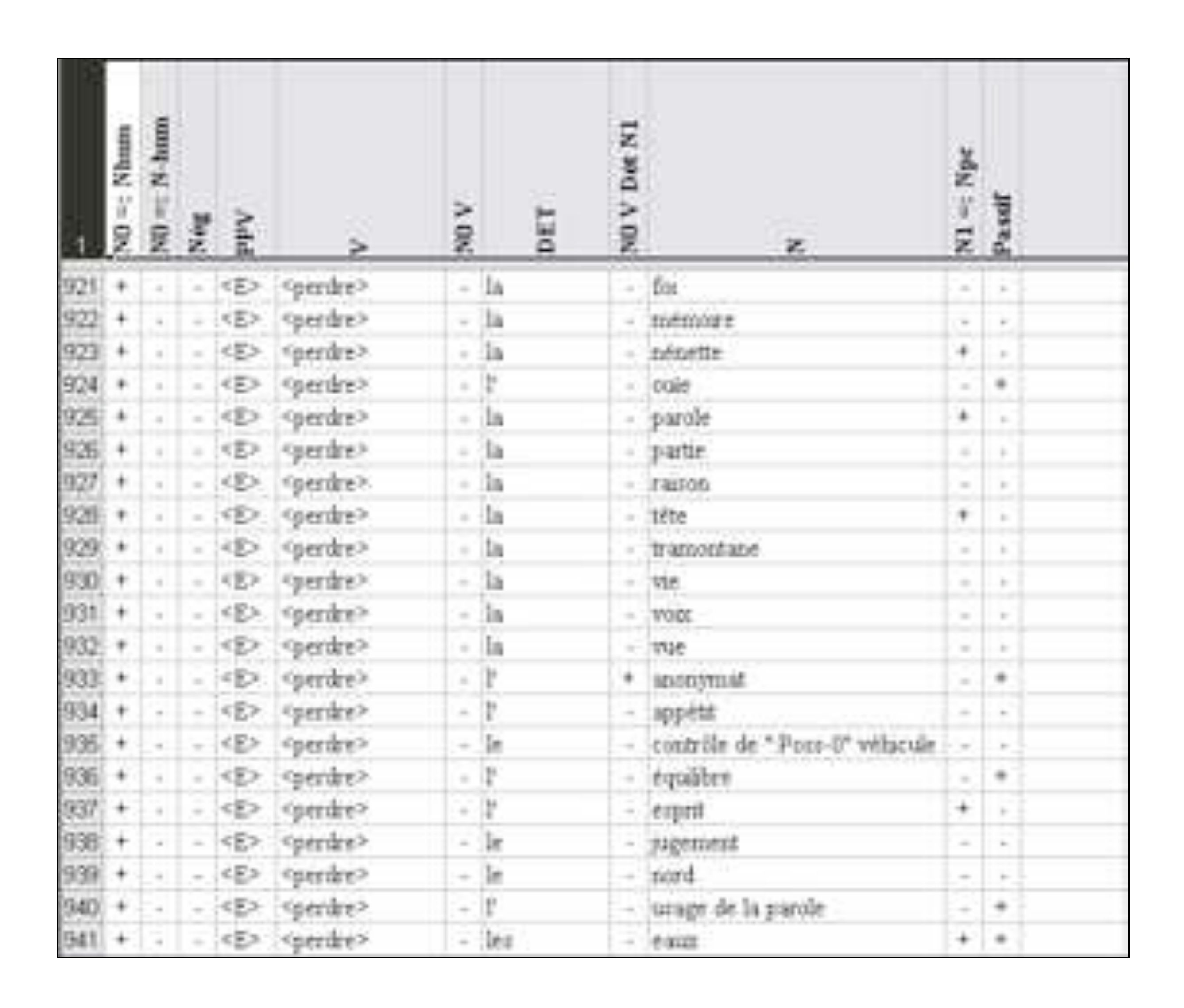
Code	Signification	Exemple
z1	langage courant	blague
z2	langage spécialisé	sépulcre
z3	langage très spécialisé	houer
Abst	abstrait	bon goût
Anl	animal	cheval de race
AnlColl	animal collectif	troupeau
Conc	concret	abbaye
ConcColl	concret collectif	décombres
Hum	humain	diplomate
HumColl	humain collectif	vieille garde
t	verbe transitif	foudroyer
i	verbe intransitif	fraterniser
en	particule pré-verbale (PPV) obligatoire	en imposer
se	verbe pronominal	se marier
ne	verbe à négation obligatoire	ne pas cesser de

Codes sémantiques

Code	Signification
m	masculin
f	féminin
n	neutre
s	singulier
p	pluriel
1, 2, 3	1ère, 2ème, 3ème personne
P	présent de l'indicatif
I	imparfait de l'indicatif
S	présent du subjonctif
T	imparfait du subjonctif
Y	présent de l'impératif
C	présent du conditionnel
J	passé simple
W	infinitif
G	participe présent
K	participe passé
F	futur

Codes flexionnels

Echantillon d'une table du Lexique-Grammaire



1	N0 =: Nhum	N0 =: N-hum	Neg	PPV	V	N0 V	DET	N0 V Det N1	N	N1 =: Npc	Passif
921	+	-	-	<E>	<perdre>	-	la	-	foi	-	-
922	+	-	-	<E>	<perdre>	-	la	-	mémoire	-	-
923	+	-	-	<E>	<perdre>	-	la	-	nénette	+	-
924	+	-	-	<E>	<perdre>	-	l'	-	ouie	-	+
925	+	-	-	<E>	<perdre>	-	la	-	parole	+	-
926	+	-	-	<E>	<perdre>	-	la	-	partie	-	-
927	+	-	-	<E>	<perdre>	-	la	-	raison	-	-
928	+	-	-	<E>	<perdre>	-	la	-	tête	+	-
929	+	-	-	<E>	<perdre>	-	la	-	tramontane	-	-
930	+	-	-	<E>	<perdre>	-	la	-	vie	-	-
931	+	-	-	<E>	<perdre>	-	la	-	voix	-	-
932	+	-	-	<E>	<perdre>	-	la	-	vue	-	-
933	+	-	-	<E>	<perdre>	-	l'	+	anonymat	-	+
934	+	-	-	<E>	<perdre>	-	l'	-	appétit	-	-
935	+	-	-	<E>	<perdre>	-	le	-	contrôle de * Poss-0" véhicule	-	-
936	+	-	-	<E>	<perdre>	-	l'	-	équilibre	-	+
937	+	-	-	<E>	<perdre>	-	l'	-	esprit	+	-
938	+	-	-	<E>	<perdre>	-	le	-	jugement	-	-
939	+	-	-	<E>	<perdre>	-	le	-	nord	-	-
940	+	-	-	<E>	<perdre>	-	l'	-	usage de la parole	-	+
941	+	-	-	<E>	<perdre>	-	les	-	eaux	+	+